



Genomic, structural and functional characterization of odorant binding proteins in olfaction of mosquitoes involved in infectious disease transmission

Malini Manoharan

► To cite this version:

Malini Manoharan. Genomic, structural and functional characterization of odorant binding proteins in olfaction of mosquitoes involved in infectious disease transmission. Agricultural sciences. Université de la Réunion, 2011. English. NNT : 2011LARE0022 . tel-00979587

HAL Id: tel-00979587

<https://theses.hal.science/tel-00979587>

Submitted on 16 Apr 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Genomic, structural and functional characterization of odorant binding proteins in olfaction of mosquitoes involved in infectious disease transmission

Thesis submitted for the
joint

Doctor of Philosophy in Life Sciences
from *Manipal University*

and

Doctorat en Sciences *spécialité* Biologie Informatique
from *Université de La Réunion*

by

Malini Manoharan

from

Laboratoire DSIMB, INSERM UMR-S 665,
Université de La Réunion
(Saint Denis, La Réunion, France)

and

Computational Approaches to Protein Sciences Group
National Center for Biological Sciences
(Bangalore, India)

on

August 22, 2011



Caractérisation génomique, structurale et fonctionnelle des protéines liant les molécules odorantes dans le système olfactif des moustiques vecteurs de maladies infectieuses

Thèse présentée pour l'obtention
conjointe du

Doctorat en Philosophie en Sciences Biologiques
de l'*Université de Manipal* (Inde)

et du

Doctorat en Sciences *spécialité* Biologie Informatique
de l'*Université de La Réunion* (France)

par

Malini Manoharan

du

Laboratoire DSIMB, INSERM UMR-S 665,
Université de La Réunion
(Saint Denis, La Réunion, France)

et du

Computational Approaches to Protein Sciences Group
National Center for Biological Sciences
(Bangalore, India)

le

22 août 2011

Certificate

We certify that this thesis entitled “Genomic, structural and functional characterization of odorant binding proteins in the olfaction of mosquitoes involved in infectious disease transmission” is an original bonafide research work of Malini Manoharan carried out under the Supervision of Prof. Sowdhamini at the National Centre for Biological Sciences and Dr. Bernard Offmann at the Université de La Reunion during the years of 2008 - 2011 for the degree of Doctor of Philosophy offered by the Manipal University and Université de La Reunion. The results described in this thesis have not been formed the basis for the award of any other degree, diploma, associateship or similar title to any university or institution.

R. Sowdhamini

Prof. R. Sowdhamini

(Thesis Supervisor)

National Centre for biological Sciences

Tata Insitute of Fundamental Research

GKVK, Bellary Road,

Bangalore 560065

India



Dr. Bernard Offmann

(Thesis Supervisor)

Laboratoire DSIMB, INSERM UMR 665

Université de La Reunion

15, avenue Rene Cassin, 97715

La Reunion

France

Declaration

I hereby declare that that the work reported in this thesis is original and was carried out by me as a PhD student under the supervision of Prof. Ramanathan Sowdhamini at the National Centre for Biological Sciences, Bangalore, India and Dr. Bernard Offmann at the Laboratoire DSIMB, INSERM UMR-S 665, Université de La Reunion, St.Denis, La Reunion, France. This thesis has not formed the basis for the award of any other degree, diploma, associateship or similar title to any university or institution.



Malini Manoharan

(22nd August , 2011)

Acknowledgements

“Gratitude is not an attitude but it is something that flows out of you when you are overwhelmed by something or somebody” -Sadhguru.

Every small progress that I have made in my life would not have been possible without the ‘immeasurable giving’ of many wonderful people around me. Apart from merely expressing it on this page, I owe them my gratitude till my last breath.

I would like to acknowledge Prof. Sowdhamini for believing in my skills as a student and providing me this opportunity to elevate myself on the topic of my research. I thank Dr. Offmann for welcoming me and supporting me at all levels, making me feel comfortable and guiding me to stay in focus on the thesis. Their professional skills have always been unbeatable and they have shaped every part of this thesis and work beautifully. They have appreciated my every small progress, respected my ideas and have always been patient with me. The expression of their enthusiasm, politeness and constant encouragement is something that I would remember forever. Apart from guiding me professionally they have also helped me shape my personality in many ways. They are wonderful human beings filled with immense beautiful qualities which have always made me fall in love with them. They are my role models in life and I enjoyed every single day of working with them. If I were to miss somebody after the completion of this work it would be them. They have been wonderful mentors than just being scientific advisors.

I would also like to extend my thanks to my thesis committee members Prof. Srinivasan and Prof. Mathew for their scintillating ideas in this work. My thesis committee meetings with them have always been fruitful and helped me look into many different dimensions of the work. I would also like to thank Dr. Patrick Fuchs for sharing his excellent knowledge on molecular dynamic simulations and guiding me through some of my work.

My lab mates at both the laboratories were wonderful in company and at work. Sandhya has always been the best person to go to when I was deprived of motivation or even for a casual talk. I have adapted a lot of her much admired presentation skills. Swati was the most frequent visitor to my desk and my best companion when it comes to fun. I would never forget the days of traveling together with Sunitha where we shared many personal and professional thoughts. I would like to thank Adwait, Prashant and Swapnil-the ‘kids’, that's what I call them even if they don't want me to. They have been there to help me in anything that I need both personally and professionally. I should mention a special note of thanks to Jayanthi and Kannan with whom I had associated to work with on certain aspects of this thesis. Though we shared the details of the work virtually, they have

proven their efficiency. I would also like to thank Gowri, Eshita, Naseer, Sony and Nagarathnam for many useful conversations. I also extend my thanks to all my other past and current lab members. Lab members in Reunion will be dear to my heart always. It just took a week for me and Magali to become good friends. She has been very supportive and helpful ever since I met her. I have admired her focus on work and the level of maturity in her. Etienne, Matthieu and Aurore have been very good associates to work with. Though young in age they have very good hearts and are always high on energy levels. I should mention that some of their questions had given out some interesting results in this work. I appreciate their sincerity, dedication and intensity towards work. I would also like to appreciate some interesting conversations on MD simulations with Lilian. I also extend my thanks to the lab members of DSIMB in Paris for their very useful discussions during my visit.

I would like to thank the Conseil Regional de La Réunion for the PhD scholarship, NCBS, University de la Reunion and Manipal university for their extensive administrative support and infrastructure. I would like to make special thanks to the computer staff in NCBS and Mme. Delphine Ramalingom in University de La reunion for helping me out with technical issues on clusters.

My stay in Reunion was made the most memorable by some families that I should mention here without whom I would have lacked a lot of moral support. I would proudly say that I have always been considered a part of the Offmann, Grondin and Dijoux family. I would just visit them whenever I wanted home and I was always taken care off at the best. The love of the kids Mark, Paul, Marion, Maurine and Anais would always remain fresh in my memories forever. I would also like to add Jayashree to these kids whose company I have enjoyed during her visits to Reunion.

I would not be able to thank my Mom, Dad and Sister just by saying thank you for anything in my life would be meaningless without them. They have always given me the best and they have dedicated their lives for my wellbeing. Every progress and every small success in my life belongs to them.

I would like to appreciate my Husband for his understanding, motivation and care he showed in building a good career for me. He was always there to support me. I owe my special thanks to my brother Anand for his motivation and support. This acknowledgement would not be complete without thanking all my awesome friends. The list is too long to mention here but I am never complete without them.

Finally, I would like to thank Sadhguru for making all this wonderful people happen to me. Nothing would have been possible without his grace for he taught me how to work joyfully.

To my MOM and DAD

Synopsis

Mosquitoes seeking their hosts or mates are exposed to a wide variety of visual, olfactory, gustatory and physical stimuli. Any one or combinations of these preferentially act as cues for host or partner identification and location (Cork et al. 1996). The role of olfaction, however, is currently found to be the major source of this identification among the mosquitoes. The molecular basis of this chemical signal recognition is systematically encoded by a series of proteins. The three major constituents involved in the peri-receptor events include the odorant binding proteins (OBPs), the odorant degrading enzymes (ODE) and the olfactory receptors (ORs) of the sensory neurons (Vogt and Riddiford. 1981). Odorant binding proteins are thought to be the primary proteins involved in the transport of odorants and pheromones to the olfactory receptors (Pelosi et al. 1995; Vogt et al. 1999). In fact, the discovery of the members in this class preceded the identification of the olfactory receptors in insects (Vogt and Riddiford. 1981). Members of this protein family have been identified in a number of insect species, including four dipterian species *Drosophila melanogaster* (Hekmat-Scafe et al. 2002; Zhou et al. 2004), *Anopheles gambiae* (Xu et al. 2003; Zhou et al. 2004), *Aedes aegypti* (Zhou et al. 2008) and *Culex quinquefasciatus* (Pelletier et al. 2009). The current research work entitled “*Genomic, structural and functional characterization of odorant binding proteins in olfaction of mosquitoes involved in infectious disease transmission*” portrayed in this thesis is focused on further characterization of the odorant binding protein family members in the genomes of *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* in their sequence, structural and functional dimensions. The thesis is organized into seven chapters explaining the methodology of investigation, the results obtained and discusses how this work opens new dimensions to the current knowledge available on mosquito odorant binding proteins.

Chapter 1 provides an overall picture of the knowledge available on odorant binding proteins family in the *Diptera* genus and presents the standing requirement for the need of its analysis from a global perspective. It provides information on global problems that drive this research, narrowing down to the importance of small proteins in a cell and their need to be studied. Computational approaches to protein science which stand as powerful tools for addressing the various questions raised in this thesis have also been described in this chapter.

Chapter 2 focuses on a genome-wide and comparative analysis of odorant binding proteins in three mosquito genomes *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*. It describes (i) the identification and extension of OBPs in these three mosquito genomes, (ii) the phylogenetic analysis of these proteins within each genome and (iii) a comparative analysis of the

different classes of these proteins between the three genomes. The results indicate a significant extension of the OBP gene family to a total of 83 new members in the three genomes. Identification of *Plus C* and *Atypical* members of the *Culex quinquefasciatus* genome and an expansion of their *Classic* OBP members, in addition to those identified by Pelletier and Leal. (2009), have been reported. The existing dataset of *A. gambiae* (Xu et al. 2003; Zhou et al. 2004) and *A. aegypti* (Zhou et al. 2008) have been enriched with new entries identified by our method. New classes of OBPs in mosquito genomes such as *Minus-C* OBPs, closely related to the *Drosophila Minus C* OBPs (Hekmat-Safe et al. 2002), and additional true *Minus -C* proteins, which lack C2 and C5 cysteines homologous to *Bombyx mori Minus C* proteins, are being described. The characteristics of odorant binding protein subfamilies in each of the genomes, using structure-based alignments and phylogeny have been highlighted, resulting in a further sub-classification of the different classes of OBPs into various subtypes. A new dimension of looking at a particular class of OBP currently described as ‘Atypical’ OBP to be dimer OBPs is provided stimulating the curiosity of the functional role of these proteins in olfaction.

Chapter 3 describes a novel method developed to identify and classify odorant binding proteins from genomic data. The method that acts a classifier of OBPs based on the cysteine conservation profiles. This involves the creation of class-specific alignment profiles carrying the cysteine conservation information and a sequence to profile alignment of queries followed by a scoring function to classify an unknown sequence. The algorithm was extended to another disulphide-rich family namely the conotoxins to show the applicability of the method to any family of proteins that can be classified on the basis of cysteine motifs and disulphide connectivity patterns. The accuracy of the method was found to be 93% and 90% for the conotoxin and OBP family respectively, proving it to be an efficient classifier of disulphide rich superfamilies. Another scoring scheme was designed especially for the OBP family based on the conservation of functionally important residues for assessing the conservation of these residues across this family of proteins.

Chapter 4 is focused on a large scale 3D-modeling of all the *classic* odorant binding proteins from the three mosquito genomes which are further used to address the functional aspects of this family of proteins in Chapter 6. A total of 135 structural models have been constructed for all classic OBPs in the three genomes. The method was based on a rigorous modeling approach that addressed the inherent divergence of the members in this class which featured low sequence identities. The alignments used for the construction of the models were obtained from consensus

fold prediction methods providing more reliable alignments based on the overall fold. These models were based as a platform for further analysis of this family in terms of function.

Chapter 5 investigates the ligand binding and release mechanism of OBPs based on molecular dynamics simulation experiments on one of the available structural members in this family under different pH conditions. It provides a description of the pH-dependent conformational adaptation of the odorant binding proteins and ligand binding states, as observed in the various molecular dynamic simulations experiments. An in-depth overview of a cascade mechanism involved in the varied conformational state of the OBP at a low pH condition is being described solving the long kindling hypothesis on the ligand binding and release mechanism.

Chapter 6 is focused on the functional aspects of the OBP family of proteins where the question of the specificity of the proteins to various ligands is investigated. It describes the functional characterization of the odorant binding proteins based on large-scale docking experiments of 135 proteins with 126 ligands and analysis of a huge dataset of 1,654,380 docked conformations to address important questions on the specificity of the OBPs..

Chapter 7 provides overall conclusions drawn from the different chapters and provides a cursory view of possible future work that can stem out of the results described in this research. As a follow up, a number of novel interesting questions about this very interesting family of proteins are being thrown open to the scientific community.

Summary

1. Introduction	1
1.1. Infectious and tropical diseases: feeling the bite of global warming.....	1
1.2. Olfaction in insects.....	2
1.2.1. Olfactory system in mosquito.....	3
1.2.2. Molecular basis of the olfactory mechanism.....	4
1.3. Odorant binding proteins.....	4
1.3.1. Structure of the odorant binding proteins.....	5
1.3.2. Mechanism of olfaction in insects involving odorant binding proteins	7
1.4. Computational approaches to protein science	9
1.4.1. Sequence searches	9
1.4.2. Phylogenetic reconstruction	10
1.4.3. Comparative modelling.....	12
1.4.4. Molecular docking.....	16
1.4.5. Molecular dynamics	19
2. Genomic characterization of odorant binding proteins in three mosquito genomes	27
2.1. Introduction.....	27
2.2. Materials and methods.....	30
2.2.1. Sequence searches	30
2.2.2. Multiple sequence alignment	31
2.2.3. Phylogenetic analysis	31
2.2.4. Chromosomal mapping.....	32
2.3. Results.....	32
2.3.1. Naming of OBP genes from mosquitoes needs to be clarified.....	32
2.3.2. Extension of odorant binding proteins family in all 3 mosquito genomes.....	34
2.3.3. Alignment of OBP proteins and description of their key sequence features.....	35
2.3.4. Analysis of OBP genes orthology across the 3 genomes and their corresponding distribution.....	37
2.3.5. Phylogenetic analysis of the odorant binding proteins.....	38
2.3.6. Comparative analysis of the Classic and Plus C subfamilies of OBPs	39
2.3.7. Sequence specific clustering of Atypical odorant binding proteins	40
2.4. Discussion	41
2.4.1. Rapid evolutionary based duplication in the Culicinae family of mosquitoes.....	41
2.4.2. Functional sub clustering of the odorant binding proteins	42
2.4.3. Atypical OBPs are indeed Two-Domain OBPs.....	43
2.4.4. Minus C proteins in the mosquito genomes.	44
2.5. Conclusion	45

3. Association of putative members to family of mosquito odorant binding proteins: scoring scheme using fuzzy functional templates and cysteine residue positions.....	60
3.1. Introduction.....	60
3.2. Methodology.....	62
3.2.1. Datasets	62
3.2.2. Construction of Profiles	63
3.2.3. Construction of fuzzy functional template.....	63
3.2.4. Scoring of query sequences.....	63
3.2.5. Composite Classification Scheme.....	66
3.2.6. Re-substitution test of the cysteine based classification scheme.....	66
3.3. Results.....	66
3.3.1. Functional Sites and Fuzzy Functional Template.....	66
3.3.2. Sequence-Based Scoring scheme.....	67
3.3.3. Cysteine-based Scoring Scheme.....	68
3.3.4. Application of scoring schemes on well-known superfamily of conotoxins	69
3.4. Discussion	70
3.5. Conclusion	71
4. Comparative modeling of classic odorant binding proteins from the mosquito genomes	80
4.1. Introduction.....	80
4.2. Materials and methods.....	83
4.2.1. Retrieval of target sequences.....	84
4.2.2. Identification of template and alignments.....	84
4.2.3. Modeling and energy minimization.....	84
4.2.4. Evaluation of refined model.....	85
4.3. Results.....	85
4.3.1. Template selection and alignment.....	85
4.3.2. Model accuracy	85
4.3.3. Structure analysis of members in a subfamily	86
4.4. Discussion	87
4.5. Conclusion	88
5. Towards unravelling the molecular mechanism underlying the functioning of an OBP through molecular dynamics simulations	103
5.1. Introduction.....	103
5.2. Methodology.....	106
5.3. Results.....	107
5.3.1. pH sensing triad.....	108

5.3.2.	Loop between helix 3 and 4	109
5.3.3.	Change in interaction patterns of helix4 and helix5	110
5.3.4.	Binding pocket and movement of MOP	110
5.3.5.	Essential dynamic analysis	111
5.4.	Discussion	111
5.4.1.	CquiOBP1 undergoes a pH dependent conformational change	111
5.4.2.	Does the previously hypothesized “pH sensing triad” of the C-terminal carboxylate contribute to conformational changes seen in the case of low pH simulations?	112
5.4.3.	Loop3 of CquiOBP1 undergoes a major conformational change	113
5.4.4.	Concerted change in interaction patterns following the conformational change of the loop	114
5.4.5.	Hypothesized exit of the ligand	114
5.5.	Conclusion	115
6.	Protein-ligand interaction profiles of Classic odorant binding proteins in the mosquito genome using molecular docking	129
6.1.	Introduction	129
6.2.	Materials and methods	131
6.2.1.	Construction of the ligand database	131
6.2.2.	Docking	131
6.2.3.	Estimation of significant interactions	132
6.3.	Results	133
6.3.1.	Optimization and validation of docking protocol	133
6.3.2.	Docking	134
6.3.3.	Analysis of the binding efficiency	134
6.3.4.	General overview on binding	135
6.3.5.	OBP binding profiles	136
6.3.6.	Comparison of the computational docking complexes with experimental docking complexes	136
6.3.7.	Characterization of binding site for known experimentally proven ligands of mosquito OBPs	137
6.4.	Discussion	137
6.4.1.	Optimization and validation of docking protocol	137
6.4.2.	SILE is a good measure for a size independent representation of the data	138
6.4.3.	Variations observed in the binding profile of odorants to ORs and OBPs - suggesting the importance of the role of OBPs for certain ligands	139
6.4.4.	Combinatorial binding profiles observed for OBPs	139
6.4.5.	Comparision of the docked complexes to experimental data	140
6.5.	Conclusion	141
7.	Conclusion	164
8.	References	173

List of Figures

	Page
Figure 1.1. Forecasted impact of climate change on the worldwide spread of malaria in 2050.	21
Figure 1.2. Estimated baseline population at risk for dengue infection in 1990 (A) and in 2085 (B) based on modelling using climate data for 1961–1990 and projections for humidity change a function of climate change—for 2080–2100.	22
Figure 1.3. Attack rates for chikungunya infections per 1,00,000 inhabitants, by administrative commune, Réunion, January 2006.	23
Figure 1.4. Details of the sensory organs and tissues that are components of the olfactory system in mosquitoes.	24
Figure 1.5. Structure of CquiOBP1 dimer from <i>Culex quinquefasciatus</i> bound to ‘3OG: (1S)-1-[(2R)-6-oxotetrahydro-2H-pyran-2-yl]undecyl acetate’ (colored in red).	25
Figure 1.6. Mechanisms by which odorants are detected in <i>Drosophila</i> .	26
Figure 2.1. Cysteine conservation patterns across the different subfamilies and subgroups of odorant binding proteins from <i>Anopheles gambiae</i> , <i>Aedes aegypti</i> and <i>Culex quinquefasciatus</i> .	46
Figure 2.2. Residue conservation patterns within each OBP subfamily from <i>Anopheles gambiae</i> , <i>Aedes aegypti</i> and <i>Culex quinquefasciatus</i> in the form of sequence logos.	47
Figure 2.3. Analysis of orthologous OBP genes shared across three mosquito species, <i>Anopheles gambiae</i> , <i>Aedes aegypti</i> and <i>Culex quinquefasciatus</i> .	48
Figure 2.4a. Rooted phylogenetic tree of the odorant binding proteins in the <i>Anopheles gambiae</i> genome.	49
Figure 2.4b. Rooted phylogenetic tree of the odorant binding proteins in the <i>Aedes aegypti</i> genome.	50
Figure 2.4c. Rooted phylogenetic tree of the odorant binding proteins in the <i>Culex quinquefasciatus</i> genome.	51
Figure 2.5a. Unrooted phylogenetic tree of <i>Classic</i> odorant binding proteins in the three mosquito genomes and in <i>Drosophila melanogaster</i> .	52
Figure 2.5b. Unrooted phylogenetic tree of <i>PlusC</i> odorant binding proteins in the three mosquito genomes.	53
Figure 2.5c. Unrooted phylogenetic tree of <i>Atypical</i> odorant binding proteins in the three mosquito genomes.	54
Figure 3.1. Alignment of available structures of odorant binding proteins using COMPARE.	72
Figure 3.2. Fuzzy functional template investigated to score the dissimilarity between OBPs.	73
Figure 3.3. Schematic representation of the investigated cysteine based scoring scheme.	74
Figure 3.4. Flowchart of the logistics used in the composite classification scheme of OBPs.	75
Figure 3.5. Flowchart of the logistics used in the composite classification scheme of the conotoxin family.	75

	Page
Figure 3.6. Effect of sequence identity on sequence based scoring scheme.	76
Figure 3.7. Effect of sequence identity on structure based scoring scheme.	76
Figure 3.8. Results of the classification schemes.	77
Figure 3.9. Cysteine connectivity patterns in the four major superfamilies of conotoxins.	77
Figure 3.10. Histogram of the number of sequences versus the % identity of the query sequence with the template.	78
Figure 4.1. Distribution of percentage identity of the OBP sequences with their respective structure template used for modelling .	89
Figure 4.2. Quality assessment of modelled OBPs as a function of sequence identity.	89
Figure 4.3. Graphical representation from PyMOL of all the <i>Classic</i> OBP models that were constructed using MODELLER.	90
Figure 4.4. Graphical representation of the superposition of models for every cluster of the <i>Classic</i> OBPs.	95
Figure 5.1. Hypothetical model for pheromone release at receptor.	116
Figure 5.2. Analysis of the putative pH sensing triad between His23, Tyr54 and Val125 in CquiOBP1.	117
Figure 5.3. RMSD plots obtained after the least mean square fit of residues 11- 124 to C-alpha atoms of starting structure of the different simulation systems.	118
Figure 5.4. RMS fluctuation plots for C-alpha atoms of every residue in the different simulation systems	119
Figure 5.5. Distances for the residues involved in the hydrogen bond triad involving the C-terminus of CquiOBP1	120
Figure 5.6. Analysis of disruption and formation of salt bridges during molecular dynamics of CquiOBP1 at pH 4.0 and pH8.0.	121
Figure 5.7. Schematic representation of the change in the conformational state of loop3 during molecular simulation at pH 4.0.	122
Figure 5.8. Analysis of hydrogen bond swapping associated with the change in the conformational state of the loop 3 during molecular dynamics of CquiOBP1 at pH 4.0.	123
Figure 5.9. Change in the interaction pattern within and between helix4 and helix5 of CquiOBP1 during molecular simulation at pH 4.0.	124
Figure 5.10. Schematic representation of the change in the interaction pattern of helix4 of CquiOBP1 during molecular simulation at pH 4.0.	125
Figure 5.11. Analysis of the dynamics of the MOP ligand and its binding cavity in CquiOBP1 during molecular simulation at pH 4.0.	126
Figure 5.12. Essential dynamics analysis of CquiOBP1 at pH 4.0.	127
Figure 6.1. Structural superposition of MOP bound to AgamOBP1 as in the crystal structure and the predicted complex	142
Figure 6.2. The docking free energies of 130 proteins against 126 ligands represented as a heat map.	142

	Page
Figure 6.3. Various plots constructed on arriving at a size independent measure of the data.	143
Figure 6.4. Heat map plot of SILE values representing the binding affinity of all the proteins against the various ligands.	144
Figure 6.5. Odorant tuning curves of ligands which indicate low binding efficiency to OBPs.	145
Figure 6.6. Ligands which indicate a broad spectrum of binding to OBPs.	147
Figure 6.7. Tuning curves of the ligands which show specific binding to the OBPs.	149
Figure 6.8. Box plot of the SILE values for broad spectrum ligands which show high binding affinity to OBPs in all the clusters.	150
Figure 6.9. Box plot of the SILE values for the repellent permethrin which shows the highest binding affinity to all the clusters.	151
Figure 6.10. Structural superposition on DEET bound to AgamOBP1 in the crystal structure and the predicted complex using AUTODOCK	152
Figure 6.11. Structural superposition on indole bound to AgamOBP4 in the crystal structure and the predicted complex using AUTODOCK.	152
Figure 6.12. Analysis of the AUTODOCK results for a set of ligands that have been shown experimentally to bind AaegOBP1, AgamOBP1, AgamOBP4, AgamOBP19 and CquiOBP1.	153
Figure 7.1. Representation of the various hypotheses about the molecular mechanism by which OBPs would be involved in olfaction in the mosquitoes.	170

List of Tables

Table 2.1. Identification of OBPs in <i>Anopheles gambiae</i> , <i>Aedes aegypti</i> and <i>Culex quinquefasciatus</i> genomes.	55
Table 2.2. Analysis of the two putative OBP domains (N-term and C-term) of <i>Atypical</i> OBPs from <i>An. gambiae</i> , <i>Ae. aegypti</i> and <i>C. quinquefasciatus</i> .	56
Table 3.1. Datasets used as training and test sets to build and assess scorings schemes for the identification of OBPs.	79
Table 4.1. List of all the <i>Classic</i> OBP models built using MODELLER.	99
Table 5.1. Description of the various time and pH conditions of the various simulations performed on CquiOBP1 (0PDB:3OGN) in this study.	128
Table 5.2. Salt bridge and hydrogen bond interactions of loop3 in the crystal structure of CquiOBP1.	128
Table 6.1. Characteristics of the odorant molecules used in this study and overview of the results of the docking experiments performed using AutoDock .	156
Table 6.2. Description of the various parameters of docking used in the optimization protocol based on the ability of AUTODOCK to reproduce the bound complex of the CquiOBP1-3OG complex for the large scale docking.	162
Table 6.3. SILE values derived from AUTODOCK energy values for ligands proved to experimentally bind to OBPs	163

List of Abbreviations

OBP: -	Odorant binding protein
OR:-	Olfactory receptor
PBP:-	Pheromone binding protein
His:-	Histidine
Lys:-	Lysine
Asp:-	Aspartate
Val:-	Valine
Gln:-	Glutamine
Phe:-	Phenyl alanine
OLDPROB:-	Old probability
NEWPROB:-	New Probability
PDB:-	Protein Data Bank
NMR:-	Nuclear Magnetic resonance spectroscopy
PAM:-	Point Accepted Mutation
BLAST:-	Basic local alignment search tool
Psi-BLAST:-	Position specific iterated blast
MEGA:-	Molecular evolutionary genetics analysis
RMSD:-	Root Mean square deviation
GA:-	Genetic Algorithm
LGA:-	Lamarckian genetic Algorithm
HA:-	Heavy atoms
FEB:-	Free energy binding
SILE:-	Size independent ligand specificity
LE:-	Ligand efficiency
FQ:-	Corrected ligand efficiency

1

Introduction

1.1. Infectious and tropical diseases: feeling the bite of global warming

Global warming refers to an increase in the average temperature of the earth, which has risen about 1°F over the past 100 years. At the current rate, the global average temperature of the earth is projected to rise from, 1.0-3.5°C by 2100 (Watson 1996), and is expected to get even warmer. As the climate continues to warm, changes are expected to occur and many effects will become pronounced over time. A spike in deadly infectious diseases may be the most immediate consequence of global warming observed at the extremes of the range of temperatures at which the transmission occurs. For many diseases these lie in the range of 14-18°C at the lower end and 35-40°C at the upper end. There has been a resurgence and redistribution of diseases like malaria and dengue vectored by the mosquitoes (Dietz et al. 1996). The ecology, development, behavior and survival of the mosquitoes and the transmission of diseases are strongly influenced by climate factors. They are sensitive to temperature changes at immature stages in the aquatic environment and as adults. The development of the larvae is sensitive to the temperature of the water and with higher temperature the time of maturity is reduced which in turn can increase the number of offsprings (Rueda et al. 1990). The digestion of a blood meal by the female mosquitoes (Gilles 1954) and the incubation period of the malaria parasites and the other viruses are also greatly influenced by increase in temperature.

According to the world malaria report, malaria is prevalent in 108 countries, with 98.5% of the deaths centering in Africa. The disease is caused by the five *Plasmodia* species (*P. falciparum*, *P. vivax*, *P. malariae*, *P. ovale* and *P. knowlesi*) among which the *P. falciparum* and *P. vivax* contribute to the significant majority of deaths vectored by *Anopheles gambiae*. The distribution of malaria is predicted to spread into new areas with temperatures suitable for the parasite *P. falciparum* (Figure 1.1). Dengue, also a potential lethal disease, was first recognized during the 1950 epidemic of Philippines and Thailand of the disease with a current incidence of disease being 2.5 billion people endemic in 100 countries. The spread of the disease is attributed to the expanding

geographic distribution of the four dengue viruses and their vector *Aedes aegypti*. Empirical models based on vapour pressure suggest that in 2085 there would be a risk of dengue transmission among 5-6 billion people in effect of the population and climate change projections compared to 3.5 million people in the absence of climate change (Figure 1.2) (Hales et al. 2002). Although malaria and dengue are the most feared infectious diseases around the world, Chikungunya has also sought attention globally after its outbreak in an unprecedented magnitude on several Indian ocean islands like Mauritius, Mayotte, Madagascar and Reunion Islands from 2001-2007. The most severe outbreak was in the Reunion islands where 7,70,000 people (one third of its population) were infected (Reiter et al. 2006) (Figure 1.3). During the same period in 2006, the disease also entrenched itself in India affecting 1.42 million people.

“We can’t solve problems by using the same kind of thinking we used when we created them.”

- Albert Einstein

The various strategies to combat these diseases continue to evolve. One of the preliminary strategies is to use mosquito control measures which include the control of mosquito egg-laying sites, control of mosquito larvae, control of mosquito adults and personal protection. One of the strategies of personal protection involves the use of repellants to prevent mosquito bites. The discovery of repellants was based on a mechanism of “olfaction” which refers to odor perception observed in living organisms and has been a cause for speculation and fascination over the centuries.

1.2. Olfaction in insects

But what are we to say of the great peacock and banded monk (moths), making their way to the female born in captivity? They hasten from the ends of the horizon. What do they perceive at that distance? Is it really an odor, as our physiology understands the word? I cannot bring myself to believe it.—Jean-Henri Fabre, from The Life of a Caterpillar, 1878.

Jean-Henri Fabre, a french entomologist, appreciated the sense of smell in insects 130 years ago based on his observations on a female peacock moth. It took another 85 years for the investigation of this stimulus to achieve isolation, purification and identification of active chemicals involved in the process. Designated as pheromones, they were defined as substances which are secreted to the outside by an individual and received by a second individual of the same species in which they release a specific reaction, for example a definite behavior of developmental process. Subsequently after this discovery another set of chemicals were described called kairomones,

semiochemicals emitted by an organism, which mediates infraspecific interactions in a way that benefits an individual of another species which receives it, without benefitting the emitter. Two main ecological cues are provided by kairomones; they generally either indicate a food source for the receiver or the presence of a predator, the latter of which is less common or at least less studied. The mosquitoes seeking the host are exposed to a wide variety of visual olfactory, gustatory and physical stimuli. Any one or combination of these preferentially acts as cues for host identification and location (Cork 1996). The role of olfaction, however, is currently found to be the major source of host identification among the mosquitoes. It was first 1950's it was identified that mosquitoes were attracted to robots with skins which had a temperature of 37°C, exhalation of CO₂, and the ones which wore jackets soaked in human sweat (Brown 1966). Scouting for potentially attractive compounds in the human sweat, researchers are taking a close look at the composition of the human sweat involving more than 300 different chemical compounds (kairomones) that contribute to the odor of the skin. These analyses are carried out using specialized instruments called olfactometers used to access the flight behavior and also by miniature electrodes attached to their nerves to sense electrical signals in response to an odor. This would help in narrowing down the attractants of the mosquito species. Another interesting side of understanding this aspect is based on studying the olfactory mechanism involved in identifying these attractants and the cell components involved with them.

1.2.1. Olfactory system in mosquito

The sensory organs of the mosquitoes are the antenna, maxillary palp, and the proboscis (Figure 1.4a). The feathery antennae serve more as general purpose olfactory organs responding to a wide range of odorants. The maxillary palp and proboscis are more tuned for close in odor and taste detection. The maxillary palp was found to have an array of specialized receptor cells for the detection of carbon dioxide and octanol, the key chemical signals involved in the identification of the human host (Lu et al. 2007). These sensory organs host hundreds of hair-like structures called sensilla attached to them which enclose the olfactory sensory neurons. The surface of each sensillum is covered with tiny pores, through which odorants pass and dissolve in a fluid called sensillum lymph, which bathes the sensory dendrites of the OSNs housed in a given sensillum. The sensillum lymph is produced by non neuronal support cells that also secrete a variety of proteins into this fluid (Figure 1.4b).

1.2.2. Molecular basis of the olfactory mechanism

The molecular basis of the chemical signal recognition is encoded by a series of proteins systematically. The three major constituents involved in the peri-receptor events include the odorant binding proteins (OBPs), the odorant degrading enzymes (ODE), and the olfactory receptors (ORs) of the sensory neurons (Vogt and Riddiford 1981). The insect olfactory receptors, being the most important components of the olfactory receptor family, were first identified in *Drosophila* with 60 genes encoding proteins with seven putative transmembrane domains. These proteins do not share any sequence homology with odorant receptors from vertebrates. Since this discovery of the complete repertoire of the *Drosophila melanogaster* ORs, candidate ORs have been identified from at least 12 insect species from four orders (*Coleoptera*, *Lepidoptera*, *Diptera* and *Hymenoptera*). It was first believed that these receptors could be G-protein coupled receptors, which trigger increases in the second messenger systems that ultimately open the ion channels. However, recently this idea was challenged by the reversed membrane topology of these receptors compared to the canonical G-protein-coupled receptors. Classic G-protein coupled seven transmembrane receptors have their C-terminus on the inside and N-terminus on the outside of the cell, while the fly ORs were found to have their C-terminus outside of the cell. The intriguing question of the involvement of G-proteins in the signaling cascade as in the other GPCRs was answered by the fact that the insect ORs are indeed ligand-gated ion channels composed of an odorant-binding OR subunit complexed with the ion conducting subunit encoded by Or83b in the case of flies. Interestingly, this could relate to the speed required to sample large volumes of air during flight which would require a faster response than a second messenger system that would involve a large number of biochemical steps. In the case of mosquitoes, a family of 79 OR genes have been identified in the *Anopheles gambiae*, 131 putative odorant receptors from the *Aedes aegypti* genome and 180 olfactory-related genes in the *Culex quinquefasciatus* genomes using computational approaches. Extensive research has been focussed on this class of proteins as potential targets for repellants.

1.3. Odorant binding proteins

The next set of members considered very important in this family are the odorant binding proteins (OBPs) which are thought to aid in the transport of odorants and pheromones to the receptors (Pelosi and Maida 1995; Vogt et al. 1999). In fact, the discovery of the members in this class predated the identification of the olfactory receptors in the insects (Vogt and Riddiford 1981). The first OBP in insects was identified in the giant moth *Antheraea polyphemus*, made up of 142

amino acid residues. Eventually, OBPs have been isolated and cloned from more than 40 insect species, belonging to ten different orders. The insect OBPs vary significantly from the vertebrate OBPs in terms of amino acid composition and three-dimensional structure. The most striking feature among all the OBPs is the conservation of six cysteines with specific spacing between them (Breer et al. 1990; Krieger et al. 1993; Krieger et al. 2005) which is considered to be the signature of this family. These cysteines are involved in disulphide bond formation providing structural integrity to the three-dimensional fold (Sandler et al. 2000; Tegoni et al. 2004; Zhou et al. 2008). The diversity of OBPs identified in the dipterians suggest that they are rapidly evolving genes through gene duplication (Vogt et al. 2002; Zhou et al. 2008). In these sequences, the similarity was only with reference to the six cysteine signature while the rest of the sequences were very divergent. Based on the divergence that was observed in this case, the OBPs were further classified into *Classic* OBPs (with standard six cysteine conservation), *Plus C* OBPs (Zhou et al. 2004) (with two additional cysteines and one proline), *Dimer* OBPs (with two cysteine signatures), *Minus C* OBPs (with the loss of two of the six conserved cysteines) and *Atypical* OBPs (with 9-10 cysteines and a long C-terminus) (Hekmat-Safe et al. 2002; Xu et al. 2003). Among the mosquito OBPs, currently 72 OBPs have been reported in *Anopheles gambiae* (Xu et al. 2003; Zhou et al. 2008), 64 OBPs in the *Aedes aegypti* (Zhou et al. 2008) and 53 OBPs in the *Culex quinquefasciatus* genome (Pelletier and Leal 2009). The odorant binding proteins considered the primary transport of the odorants to the olfactory receptors stated above have emerged as novel targets for repellants.

1.3.1. Structure of the odorant binding proteins

The first odorant binding protein structure studied was the *Bombyx mori* PBP both in its crystallized form, by X-ray diffraction spectroscopy (Krieger et al. 2005) and in solution using nuclear magnetic resonance (NMR) techniques (Damberger et al. 2000; Sandler et al. 2000). Since then, 62 structures in total, including ligand-bound and mutant forms, have been deciphered from different organisms (Rothmund et al. 1999; Sandler et al. 2000; Horst et al. 2001; Lee et al. 2002; Kruse et al. 2003; Lartigue et al. 2004; Mohanty et al. 2004; Lautenschlager et al. 2005; Zubkov et al. 2005; Wogulis et al. 2006; Damberger et al. 2007; Lautenschlager et al. 2007; Laughlin et al. 2008; Pesenti et al. 2008; Thode et al. 2008; Pesenti et al. 2009; Mao et al. 2010). The structure of these proteins show that they are mainly folded into alpha helices which are packed compactly due to the presence of three disulphide bonds formed between the conserved cysteines in the family described above (Figure 1.5). The first crystal structure in this family (*Bombyx mori* OBP bound to

bombykol; PDB ID: 1DQE) is represented as a rough conical structure of six helices four (1, 4, 5, 6) of which converge to form the hydrophobic binding pocket capped by helix 3. The disulphide bonds stabilize the position of helix 3 by attaching it to the flanking helices (5 and 6) and the other disulphide bond bridges 5 and 6 resulting in a rigid compact structure. The solution structure of the same protein at a pH 4.5 (PDB ID : 1GM0) referred to as the A-form BmPBP showed a striking conformational difference, described in (Horst et al. 2001), where the C-terminal segment consisting of approximately 10 residues located on the surface of the crystal structure (Sandler et al. 2000) was identified to form a helix structure located at the protein core in the ligand-free form. The same group eventually reported the NMR structure of the free form of the protein at pH 6.5 (PDB ID : 1LS8) described as B-form where the C-terminal was located out of the binding cavity providing a sufficiently large hydrophobic cavity, indicating an active conformational state for ligand binding. Subsequent crystal structure of the B-form of the protein at pH 7.5 (PDB ID : 2FJY) surprisingly showed that it was more similar to the A-form of the protein described at pH 4.5 demonstrating that at least two conformations of the protein can exist at neutral pH, and that the equilibrium between the conformations are sensitive to the presence or absence of ligand, along with pH (Lautenschlager et al. 2005). The NMR structure of Apol PBP1 at pH 4.5 from *Antheraea polyphemus* also forms a similar confirmation of the C-terminal helix occupying the binding pocket. These structures shed some light on the mechanism of ligand binding and release mechanism of odorant binding proteins further described further in this chapter.

The next best described structural members of this family are the apo-form (Thode et al. 2008), alcohol bound forms (Kruse et al. 2003) and pheromone (cVA) bound forms of OBP LUSH. The structure was similar to the *Bombyx mori* pheromone binding protein with the major difference observed in the packing of the helix 1 and the C-terminal tail. In Lush, the C terminal tail folds into the core and a part of the alcohol binding pocket, whereas helix 1 packs outside the protein whereas the reverse is observed in the BmPBP complex. This suggested an alternate mechanism for olfaction in these proteins, subsequently described in this chapter, where the conformational changes of the bound complex directly triggers the mechanism without the release of the ligand (Laughlin et al. 2008). Similar conformation of the C-terminal was also observed in *Apis mellifera* OBP (Amel-ASP1) with conformational differences observed with respect to the nature of the ligands providing flexibility to the binding site. This structure also shows the presence of a beta hairpin structure observed between helix 3 and 4 of the structure. Four odorant binding crystal structures have been described in the mosquito genomes two from *Anopheles gambiae* (AgamOBP1) (Wogulis et al. 2006); AgamOBP22 (*no citation*) and one from *Aedes aegypti* (Leite et al. 2009) and *Culex*

quinquefasciatus (Mao et al. 2010) each. The structures are closely related to the *Drosophila* Lush, with a short C-terminal buried in the binding pocket, except AgamOBP22 which shows a slightly different fold from all the above proteins; however, the citation of this PDB entry is currently not available for further description.

All the above described structures belong to the Classic OBP class of the odorant binding protein family and recently the structure of a member of the C+ class of OBP was illustrated (PDB code:3PM2; (Lagarde et al. 2011). The Plus C OBP (AgamOBP47) is made up of 173 residues long with 13 cysteines and folds into eight helices rather than six helices as observed in the classic OBP structures. Superimposition of Lush with AgamOBP47 showed that the helices which embrace the conserved OBP domain are superposed very well. Helix 2 holds different spatial location between the two proteins and the later additionally has 2 extra helices (H6 and H8). The additional cysteines form 3 additional disulphide bonds located between the N-terminal and C-terminal parts of the protein resulting in a rather flat structure compared to the classic OBPs.

1.3.2. Mechanism of olfaction in insects involving odorant binding proteins

The role of odorant binding proteins in the mechanism of olfaction is so far been best explained in the *Drosophila melanogaster* genome based experiments using Lush, an extracellular odorant binding protein located in the trichoid sensillum (Kim et al. 1998; Xu et al. 2005). This acts as the receptor for the pheromone 11 *cis* vaccenyl acetate (cVA) secreted in the males for the attraction of the female species (Bartelt et al. 1985; Xu et al. 2003). cVA triggers a conformational change in the odorant binding protein LUSH which is recognized by the receptors in the T1 neurons (Laughlin et al. 2008). This was demonstrated using a LUSH D118A mutation which mimicked the active site of Lush in the absence of cVA (Laughlin et al. 2008). The receptor for the activated LUSH comprises of three subunits OR67d belonging to the olfactory family of receptors (Ha and Smith 2006; Kurtovic et al. 2007), OR83b an ion channel and SNMP which is an homologue of CD36 (Benton et al. 2007; Collot-Teixeira et al. 2007) involved in the uptake of lipoproteins in the humans. Mutants without this SNMP protein and with both OR67d and OR83b do not respond to active lush which suggests that they physically interact with the LUSH (Jin et al. 2008). The activation of the response mediated by conformational change could correspond to the single molecule sensitivity for which the insect pheromone detection is renowned. If cVA binds to LUSH at a slow rate the activated LUSH - cVa complex should be long lived and would be free to diffuse in the sensillum until it encounters the target receptors in the membrane. However, studies indicate

that the protein receptor complex of LUSH is still incomplete and could involve additional members in this process which are yet to be deciphered. Nevertheless, it has been well described in the *Drosophila* that the olfactory mechanism is triggered by the conformational change observed in the LUSH without the need of the direct release of the ligand in this case (Figure 1.6) (Ronderos and Smith 2009).

Initial analysis of the *Bombyx mori* odorant binding protein using circular dichroism and fluorescence spectroscopy showed that its tertiary structure was sensitive to pH changes and that a dramatic conformational transition occurred between pH 5.0 and 6.0 (Wojtasek and Leal 1999). Further, NMR based conformational analysis revealed interesting features of the protein. The NMR structure at pH 4.5, in contrast to the crystal structure which was deciphered at a pH of 8.2, showed the presence of 7 helices where the extended C - terminal end in the crystal structure folds into a helix buried inside the binding pocket in replacement of the ligand Bombykol at low pH. This helix is surrounded by helices 3a, 4, 5, and 6 and covered by the loop that links helices 3a and 4. The conformation of this loop is stabilized by hydrogen bonds. First half of the loop contains a classic type II Beta turn with standard i, i+3 hydrogen bonds. Type I Beta turn is formed by residues 69-72 which cover the C-terminal end of the helix 7. The residues which interact with the helix 7 in this conformation were also involved in binding of the pheromone to the protein. The residues which make direct contact with Bombykol do not interact with helix 7. In addition, the histidines His-69 His-70 His-95, which were proposed to contribute in a pH-dependent conformational change (Sandler et al. 2000), were widely separated in the NMR structure. Such a separation could reduce the charge repulsion caused by the protonation of these residues destabilizing the BmPBP–bombykol complex. If one accepts the hypothesis that the pH near the membrane surface is lower than the pH value of 6.5 measured in the bulk sensillar lymph, then it is reasonable to speculate a rationale for destabilization of the BmPBP–bombykol complex near the membrane-standing pheromone receptor, which would lead to ejection of the ligand making it available to the receptor.

Assuming that this conformational transition could correlate to the ligand-binding release pathways, a later replica exchange molecular dynamics study (Grater et al. 2006) proposed two opposite dissociation routes which could serve as the entrance/exit of the ligand. The first passage was along the front lid formed by residues 60-68 and the second one located close to the N and C terminal of the protein. These two regions were found to be highly flexible, forces and free energy calculations also revealed that both the pathways were physiologically relevant.

This thesis describes the efforts in analyzing the above stated Odorant Binding Proteins (OBPs) in three Mosquito genomes *Anopheles gambiae*, *Aedes Aegypti* and *Culex quinquefasciatus*

using powerful computational tools described further below in this chapter. The analysis and the various results thus derived are explained in the various Chapters of this thesis.

1.4. Computational approaches to protein science

1.4.1. Sequence searches

Homology is a powerful tool that helps in the identification of functionally related proteins. This descends from the fact that functionally important residues are conserved during evolution and two homologous proteins show high functional similarity. Detection of homology is based on the likelihood of the evolution of the two sequences from a common ancestor. This can be achieved by the identification of common patterns i.e. the similarities and dissimilarities of the sequence with respect to the ancestral sequence. Dynamic programming methods aid in the alignment or matching of two sequences wherein two sequences are laid across a two-dimensional matrix and then compared with each other. The similarities and dissimilarities are represented as numerical values obtained from substitution score matrices which are associated with the likelihood of residue exchange. Different substitution matrices have been described such as Dayhoff matrix (Dayhoff et al. 1983) which is based on the examination of closely related sequences in different families of proteins, PAM matrices which account for mutation, JTT (Jones et al. 1992) and Gonnet matrices (Gonnet et al. 1992) which are derived from multiple sequence alignments and the BLOSUM matrices that are obtained from local alignments of related proteins. However, the choice of the matrix depends on the nature of the sequences under question. BLOSUM62, Gonnet and the Johnson-Overington matrices have been shown to perform well in the case of distantly related proteins (Henikoff and Henikoff 1993). The matrix is then used to trace the alignment path and recognise the path where the sum of substitution scores along the path is maximal. All possible paths are evaluated to arrive at a progressive, sequence consecutive alignment in some approaches (Needleman and Wunsch 1970), while others are more tuned to locate local sub-alignments and extend them, if feasible (Smith and Waterman 1981). To arrive at the best alignments, gaps may have to be inserted into the compared sequences to mimic evolutionary processes such as insertion or deletion which is later penalized from the final score. This method is applied in search methods like BLAST and FASTA to achieve global and local alignments respectively.

1.4.1.1. *PSI-BLAST*

Database searches using position specific score matrices, also called profiles or motifs, often are more better able to detect weak relationships than a simple query search. PSI-BLAST is considered to be the most sensitive BLAST programs and is highly useful for divergent family of proteins which retain only certain signatures while the rest of the sequence is completely unrelated. The method of PSI-BLAST involves a series of repeated steps or iterations. First, a database search of a protein sequence database is performed using a query sequence. Once a list of related sequences have been identified, the process is iterated by searching the database again using a scoring matrix that indicates the variation at each aligned position from the alignment of high scoring sequence matches found in the first run. The iteration can be continued where new alignments are created with the newly identified sequence creating a refined scoring matrix. This process is continued either till no more new sequences are identified or until a user-defined threshold is reached (Altschul et al. 1997). PSI-BLAST is available as a part of the NCBI BLAST and the offline BLAST package.

1.4.1.2. *CLUSTAL X*

CLUSTALX is a general purpose progressive multiple sequence alignment program based on dynamic programming algorithm which produces biologically meaningful multiple sequence alignments of divergent sequences. CLUSTALX helps to locate the identities, similarities and differences between sequences. In CLUSTALX individual weights are assigned to each sequence in a partial alignment in order to down-weight near-duplicate sequences and up-weight the most divergent ones, after which the amino acid substitution matrices are varied at different alignment stages according to the divergence of the sequences to be aligned. CLUSTALX introduces residue-specific gap penalties and locally-reduced gap penalties in hydrophilic regions encourage new gaps in potential loop regions rather than regular secondary structure. Finally, positions in early alignments, where gaps have been opened, receive locally reduced gap penalties to encourage the opening up of new gaps at these positions (Thompson et al. 1994).

1.4.2. **Phylogenetic reconstruction**

A phylogenetic analysis of a family of related protein or nucleic acid sequences is a determination of how the family might have been derived during evolution. They also help in prediction of functional relationships of genes or proteins in a family. The theoretical frameworks

for molecular systematics were laid in the 1960s in the works of Emile Zuckerkandl, Emanuel Margoliash, Linus Pauling and Walter M. Fitch. Within the past decade, this field has been further re-energized and re-defined as whole genome sequencing for complex organisms has become faster and less expensive. As mounds of genomic data becomes publicly available, molecular phylogenetics is continuing to grow and find new applications. Procedures for phylogenetic analysis are strongly linked to the sequence alignment. Just as two similar sequences can be aligned easily, they are also easily organized into a tree but as their divergence increases, the complexity of organizing the tree and considerable expertise is required in such situations. The most common method of multiple sequence alignment is the progressive alignment which first aligns most closely related pair of sequences and then sequentially adds more distantly related sequences to this initial alignment. Gaps in alignments can be thought of as representing mutational changes in sequences, including insertions, deletion or rearrangement of genetic material. Gaps are treated in various ways by phylogenetic programs. Some methods ignore them and in some cases they can be used as biological markers. Another approach to handle gaps is to avoid individual sites in an alignment and generate a similarity score based on a scoring matrix with penalties for gaps and converting them to distance scores that are suitable for phylogenetic analysis by distance methods.

1.4.2.1. Maximum parsimony

The maximum parsimony method predicts the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences. For each aligned positions in the alignment, phylogenetic trees that require the smaller number of evolutionary changes to produce the observed changes are identified and the tree with the smallest changes is determined. This method is best suited for sequences that are quite similar and for aligning a small number of sequences.

1.4.2.2. Distance methods

The distance method employs the number of changes between each pair in a group of sequences to produce a phylogenetic tree of the group. The sequence pairs that have the smallest number of sequence changes are termed neighbours and share a common ancestor called the node and are represented as branches connected to the node. The goal of the distance methods is to identify a tree that positions neighbours correctly and also has branch lengths to reproduce the original data. Different algorithms are used in the distance based methods and here the neighbour joining method is explained in detail.

Neighbour joining method (Saitou and Nei 1987) is especially suitable for sequences where the rate of evolution varies considerably. Neighbour joining chooses the sequences that should be

joined to give the best least square fit estimates of the branch lengths that most close reflect the actual distances. First, the distance between a pair of sequence are used to calculate the sum of the branch lengths for a tree to form a star-shaped tree which is further modified based on the branch lengths. Next, a new table with this pair of sequence as a single composite sequence is produced. This is continued for the next pair until the correctly branched tree and the branch distances have been identified.

1.4.2.3. *Maximum likelihood methods*

This method uses probability calculations to find a tree that best accounts for the variation in a set of sequences. The method is similar to the maximum parsimony method where every column in the alignments are accounted for the analysis and hence it is suitable for a small set of sequences. However, this method provides an additional opportunity to evaluate trees with variations in the mutation rates in different lineages. This method is unfortunately computational intensive but however it solves complex models of evolution.

1.4.2.4. *MEGA 4.0*

Since the early 1990s, MEGA software functionality has evolved to include the creation and exploration of sequence alignments, the estimation of sequence divergence, the reconstruction and visualization of phylogenetic trees, and the testing of molecular evolutionary hypotheses (Tamura et al. 2007). The software facilitates the construction of trees using different methods like Neighbor joining, UPGMA, Maximum parsimony and Maximum likelihood methods. It also provides a user friendly interface providing an easy access to the various options and facilitating various display options of the trees. The software is available as a native 32-bit Windows application with multi-threading and multi-user supports, and it is also available to run in a Linux desktop environment and on intel based Machintosh computers under Parallels program. It is an open source software available for download at <http://www.megasoftware.net>.

1.4.3. **Comparative modelling**

Comparative or homology protein structure modelling which builds a three-dimensional model for a protein of unknown structure (the target) based on one or more related proteins of known structure (the templates) (Blundell et al. 1987; Sali and Blundell 1993; Johnson et al. 1994; Sali 1995; Sanchez and Sali 1997a; Sanchez and Sali 1997b; Fiser et al. 2000; Marti-Renom et al.

2000; Sanchez et al. 2000; Fiser et al. 2002) is the only method that can reliably predict the structure comparable to a low resolution experimentally determined structure (Marti-Renom et al. 2000).

The necessary conditions for getting a useful model are (i) detectable similarity between the target sequence and the template structures and (ii) availability of a correct alignment between them. The comparative approach to protein structure prediction is possible because a small change in the protein sequence usually results in a small change in its 3D structure (Chothia and Lesk 1986). It is also facilitated by the fact that 3D structure of proteins from the same family are more conserved than their primary sequences (Lesk and Chothia 1980). Therefore, if similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed. Moreover, proteins that share low or even non-detectable sequence similarity many times also have similar structures.

The general steps followed in comparative modeling include :

- ✓ searching and selecting for structures related to the target sequence,
- ✓ aligning the target sequence with one or more structures,
- ✓ model building,
- ✓ evaluating a model.

1.4.3.1. Searching and selecting for structures related to the target sequence

Comparative modeling requires a suitable template that is an available structure which acts as a backbone for the unknown protein. Generally, this is done by searching for similar structures in the PDB (Protein data bank). This can be done in different ways where the target sequence is compared with a known database using sequence-sequence pairwise comparison data (Apostolico and Giancarlo 1998) with the help of programs like BLAST (Altschul et al. 1997) and FASTA (Pearson 2000). This can also be further modified to perform multiple sequence comparisons to improve sensitivity of the search (Henikoff and Henikoff 1992; Krogh et al. 1994; Gribskov and Veretnik 1996; Altschul et al. 1997; Jaroszewski et al. 1998) where PSI-BLAST (Altschul et al. 1997) is used. Another method called threading or 3D template matching (Bowie et al. 1991; Overington et al. 1992; Johnson et al. 1994; Godzik 1996) is used when there are no potential templates recognizable by a simple sequence search. A sequence identity above 30% is considered to be good for carrying out comparative modeling.

1.4.3.2. *Aligning the target sequence with one or more structures*

All available comparative modeling programs depend on the structural equivalences between the target and the template. In such cases, the alignment is relatively simple to obtain using sequence alignment programs when the target-template sequence identity is high but when the target-template sequence identity is lower than 40%, the alignment requires manual intervention to minimize the number of misaligned residues and also to remove the gaps. This can be done based on the secondary structure information where gaps can be avoided in secondary structure elements, in buried regions, or between two residues that are far in space. The editing of the alignment is considered highly important because an error of approximately 4Å could be observed in the model for every single misaligned residue (Sanchez and Sali 1997a; Blake and Cohen 2001; Jennings et al. 2001; Shi et al. 2001).

1.4.3.3. *3D Jury metaserver*

3DJury is a powerful procedure for generating meta-predictions using variable sets of models obtained from diverse sources (Ginalski et al. 2003). Owing to the fact that consensus structure prediction methods have higher accuracy than individual structure prediction algorithms, the resulting protocol should help to improve the quality of structural annotations of novel proteins. The 3DJury takes a set of models generated by a set of servers as input to compare them based on C α RMSD and assigns a similarity score. The final 3D-Jury score of a model is the sum of all similarity scores of considered model pairs divided by the number of pairs considered plus one. The 3D-Jury system neglects the confidence scores assigned to the models by the servers and is based on the expectation that highly reliable models produced by the fold recognition methods have less ambiguities in the alignments. The alignments are available to the user for further use with other structure predicting softwares. The 3D-Jury system is available via the Structure Prediction Meta Server (<http://BioInfo.PL/Meta/>) to the academic community (Ginalski et al. 2003).

1.4.3.4. *Model building*

Model building can be done by different methods:

(a) Modelling by assembly of rigid bodies

This is a semi-automated procedure implemented in the program COMPOSER (Sutcliffe et al. 1987). In this, a model is assembled using small number of rigid bodies obtained from the aligned protein structural templates (Browne et al. 1969; Blundell et al. 1987). This basically

involves the dissection of the protein into conserved secondary structures, variable loop regions and the side chains that connects them.

(b) Modeling by segment matching or coordinate reconstruction

In this method, comparative models are constructed by using a subset of atomic positions from template structures as “guiding” positions, and by identifying and assembling short, all-atom segments that fit these guiding positions (Bystroff and Baker 1998).

(c) Modeling by satisfaction of spatial restraints

This method begins by generating many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. This is based on the NMR derived restraints which are obtained by assuming that the corresponding distance between aligned residues in the template and target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints such as bond lengths, bond angles, dihedral angles and non-bonded atom-atom contacts that are obtained from a molecular mechanics force field. The model is then derived by minimizing the violations of all the restraints. This can be achieved by distance geometry or real space optimization (Sali and Blundell 1990; Sali and Blundell 1993; Sali and Overington 1994; Fiser et al. 2000). This method is implied in the most widely used modeling program MODELLER.

1.4.3.5. MODELLER

As stated earlier, MODELLER is a program used for homology modeling of protein three-dimensional structures based on the alignment provided by the user for satisfaction of spatial restraints. Distance and dihedral angle restraints on target sequence are derived from its alignment with the template 3D structure. The form of these restraints was obtained from a statistical analysis of relationships between similar protein structures. The analysis relied on a database of 105 family alignments that included 416 proteins of known 3D structure. By scanning the database of alignments, tables quantifying various correlations were obtained, such as correlation between two equivalent C^α-C^α distances or between equivalent main chain dihedral angles from two related proteins. These relations are expressed as conditional probability density functions and can be used directly as spatial restraints. Probabilities for different values of main chain dihedral angles are calculated from type of residue considered, from main chain conformation of an equivalent residue and from sequence similarity between two proteins. In the second step, the spatial restraint and

CHARMM 22 force field terms enforcing proper stereochemistry are combined into an objective function. The objective function depends on Cartesian coordinates of approximately 10,000 atoms that form the modeled molecules. The function form of each term is simple, it includes a quadratic function, harmonic lower and upper bounds, a weighted sum of a few gaussian functions, Coulomb law, Lennard-Jones potential and cubic splines. The geometric features include a distance angle, dihedral angle, a pair of dihedral angles between 2, 3, 4 and 8 atoms, shortest distance in set of distances, solvent accessibility in Å and atom density that is expressed as number of atoms around a central atom. Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of variable target function method employing methods of conjugate gradient and molecular dynamics with simulated annealing. This is considered as one of the most reliable techniques for model building.

1.4.3.6. *Evaluating a model*

The quality of a model primarily depends on the sequence similarity between the target and the template and a sequence identity of above 30% is relatively a good indicator of expected accuracy. The evaluation of a model can be two types “Internal” evaluation of self-consistency checks whether or not a model satisfies the restraints used to calculate it that is the assessment of the models stereochemistry such as bonds, bond angles, dihedral angles, and non-bonded atom-atom distances. This is widely done with the help of programs such as PROCHECK (Laskowski et al. 1996) and WHATCHECK. The external evaluation involves testing the compatibility between the sequence and structure based on Z-score and also by the prediction of unreliable regions in the model using a “pseudo energy” profile with the help of servers like PROSA (Sippl 1995). Finally, a model should be consistent with experimental observations, such as known function site information, site-directed mutagenesis, cross-linking data, and ligand binding.

1.4.4. **Molecular docking**

Generally speaking, molecular docking comprises the process of generating a model of a complex based on the known 3D structures of its components, free or complexed with other species. Pioneered during the early 1980s, it remains a field of vigorous research, having become a useful tool in drug discovery efforts, and a primary component in many drug discovery programs. In particular, protein–ligand docking occupies a very special place in the general field of docking, because of its applications in medicine (Muegge et al. 2001). From the initial efforts involving the

docking of both protein and ligand as rigid bodies, protein–ligand docking has evolved to a level where full or at least partial flexibility on the ligand is commonly employed. Over the last years, several important steps beyond this point have been given. Handling the flexibility of the protein receptor efficiently is currently considered one of the major challenges in the field of docking. The fact that proteins are in constant motion between different conformational states with similar energies is still often disregarded in docking studies, even though protein flexibility is known to allow increased affinity to be achieved between a given drug and its target (Teague 2003) . Furthermore, binding site location and binding orientation can be greatly influenced by protein flexibility.

In terms of protein–ligand docking methods, the docking problem can be rationalized as the search for the precise ligand conformations and orientations (commonly referred as posing) within a given targeted protein when the structure of the protein is known or can be estimated. The binding affinity prediction problem addresses the question of how well the ligands bind to the protein (scoring). Docking protocols can be described as a combination of a search algorithm and a scoring function. The search algorithm should allow the degrees of freedom of the protein–ligand system to be sampled sufficiently as to include the true binding modes. Naturally, the two critical elements in a search algorithm are speed and effectiveness in covering the relevant conformational space. Among other requirements, the scoring function should represent the thermodynamics of interaction of the protein–ligand system adequately as to distinguish the true binding modes from all the others explored, and to rank them accordingly. Furthermore, it should be fast enough to allow its application to a large number of potential solutions. Logically, the ideal solution would be to combine the best searching algorithm with the best scoring function. However, several studies have shown that the performance of most docking tools is highly dependent on the specific characteristics of both the binding site and the ligand to be investigated, and that establishing which method would be more suitable in a precise context is almost impossible (Charifson et al. 1999; Bissantz et al. 2000; Halperin et al. 2002)

1.4.4.1. AUTODOCK 4.0

AUTODOCK 4 is a novel and robust automated docking method that predicts the bound conformations of flexible ligands to macromolecular targets such as proteins, enzymes, antibodies, DNA and RNA in combination with a new scoring function that estimates the free energy change upon binding. AUTODOCK uses a Lamarckian genetic algorithm (LGA), but encompasses also a

Monte Carlo simulated annealing and a traditional genetic algorithm. However, the last two are not as efficient and reliable as the LGA (Morris *et al*, 1998).

Genetic algorithms (GA) employ ideas based on the language of natural genetics and biological evolution. In the case of molecular docking, the particular arrangement of a ligand and a protein can be defined by a set of values describing the translation, orientation, and conformation of the ligand with respect to the protein: these are the ligand's *state variables* and, in GA, each state variable corresponds to a gene. The ligand's state corresponds to the genotype, whereas its atomic coordinates correspond to the phenotype. In molecular docking, the *fitness* is the total interaction energy of the ligand with the protein and is evaluated using the energy function. Random pairs of individuals are mated using a process of *crossover*, in which new individuals inherit genes from either parent. In addition, some offsprings undergo random *mutation*, in which one gene changes by a random amount. *Selection* of the offspring of the current *generation* occurs based on the individual's fitness: thus, solutions better suited to their environment reproduce, whereas poorer suited ones die.

Classical genetic algorithms represent the genome as a fixed length string and employs binary crossover and binary mutation to generate new individuals in the population. These genetic algorithms are based on the characteristics of Darwinian evolution and apply Mendelian genetics. LGA is based on Lamarcks assertion that phenotypic characteristics acquired during an individual's lifetime can become heritable traits. In simple words, the xyz coordinates of each conformer are encoded as a string (say 10.20.30 for a really simple coding of one atom). Random mutation are made to the strings and crossovers between them during breeding. During the 'life' of a conformer, it may do some local movements that can be transmitted to its offspring. This phenotypic change is then recorded into its genotype. This is a combination of GA method with the adaptive LS method and it is found to have an enhanced performance than the normal simulated annealing and genetic algorithm docking as it employs both the genetic algorithm and the local search. In the case of autodock, the fitness or the energy is calculated from the ligands coordinates, which together form its phenotype.

Many of the traditional force fields model the interaction energy in terms of dispersion, repulsion, hydrogen bonding, electrostatics and deviation from ideal bond lengths and bond angles. These approaches tend to perform less well in ranking the binding free energies of compounds that differ by more than a few atoms. AUTODOCK uses an empirical binding free energy function which is calculated as

Final intermolecular energy + Final internal energy + Torsional energy –Unbound system's energy.

Final intermolecular energy represents Van der waals energy + Hydrogen bond energy + Electrostatic energy + Desolvation energy: where for Van der waals energy a Lennard Jones 12-6 dispersion repulsion term is used, for hydrogen bond energy a directional 12-10 hydrogen bonding term is used, for electrostatic energy a coulombic electrostatic potential is used and for desolvation energy (which is the most challenging model) the desolvation upon binding and the hydrophobic effect (solvent entropy changes at solute-solute interface). Internal energy represents change in the internal energy of the ligand upon binding. Torsional energy represents the restriction of internal rotors, global rotation and translation. AUTODOCK also implies an empirical free energy function to determine the binding constants. This was calibrated using 30 structurally known protein-ligand complexes with experimentally determined binding constants and was found to work efficiently.

1.4.5. Molecular dynamics

Molecular dynamics investigate the motions of a system of discrete particles under the influence of internal and external forces providing the fluctuations and conformational changes in proteins and nucleic acids. The principle behind this is that interactions of the respective particles are empirically described by a potential energy function from which the forces that act on each particle are derived. With the knowledge of these forces it is possible to calculate the dynamic behavior of the system using a classical equation of motion in their simplest form that is Newton's law. The result is a trajectory that specifies how the positions and velocities of the particles in the system varies with time obtained by solving differential equations embodied in newtons second law ($F=ma$).

1.4.5.1. GROMACS

GROMACS provides a versatile and efficient molecular dynamics program, written using in C language, especially directed towards the simulation of biological macromolecules in aqueous and membrane environments, and be able to run on single and parallel computer systems (Van Der Spoel et al. 2005). It also provides with stochastic dynamics and energy minimization in addition to the Hamiltonian mechanics. Various coupling methods to temperature and pressure bath are provided to check on the stability of the system. External forces can be applied to enforce non equilibrium dynamics or steered MD. Atom grouping is facilitated for participation and analysis of

the dynamics. It includes a range of analysis tools, starting from extensive trajectory analysis to normal mode and principal component analysis of structure fluctuations. The force fields used for the intramolecular interactions is, in principle, not part of GROMACS but facilitates the use of external force fields like AMBER (Case et al. 2005), CHARMM (Brooks et al. 2009), Coarse grained force fields, GROMOS (Walter R. P. Scott 1999) and OPLS (William L. Jorgensen 1996). These force fields are computed on the basis of three different types of interactions: bonded, non-bonded and special interactions which are the restraints of the position distance or angle.

1.4.5.2. VMD

VMD is a freely available efficient molecular graphics program designed for the visualization and analysis of molecular trajectories of proteins and nucleic acids (Humphrey et al. 1996). VMD, written in C++, can efficiently display any number of structures using a wide variety of rendering styles and coloring methods. Molecules are displayed as one or more “representations”, in which each representation embodies a particular rendering method and coloring scheme for a selected subset of atoms. The atoms displayed in each representation are chosen using an extensive atom selection syntax, which includes Boolean operators and regular expressions. It is also provided with a complete graphical user interface for program control, as well as a text interface using the Tcl embeddable parser to allow for complex scripts with variable substitution, control loops, and function calls.

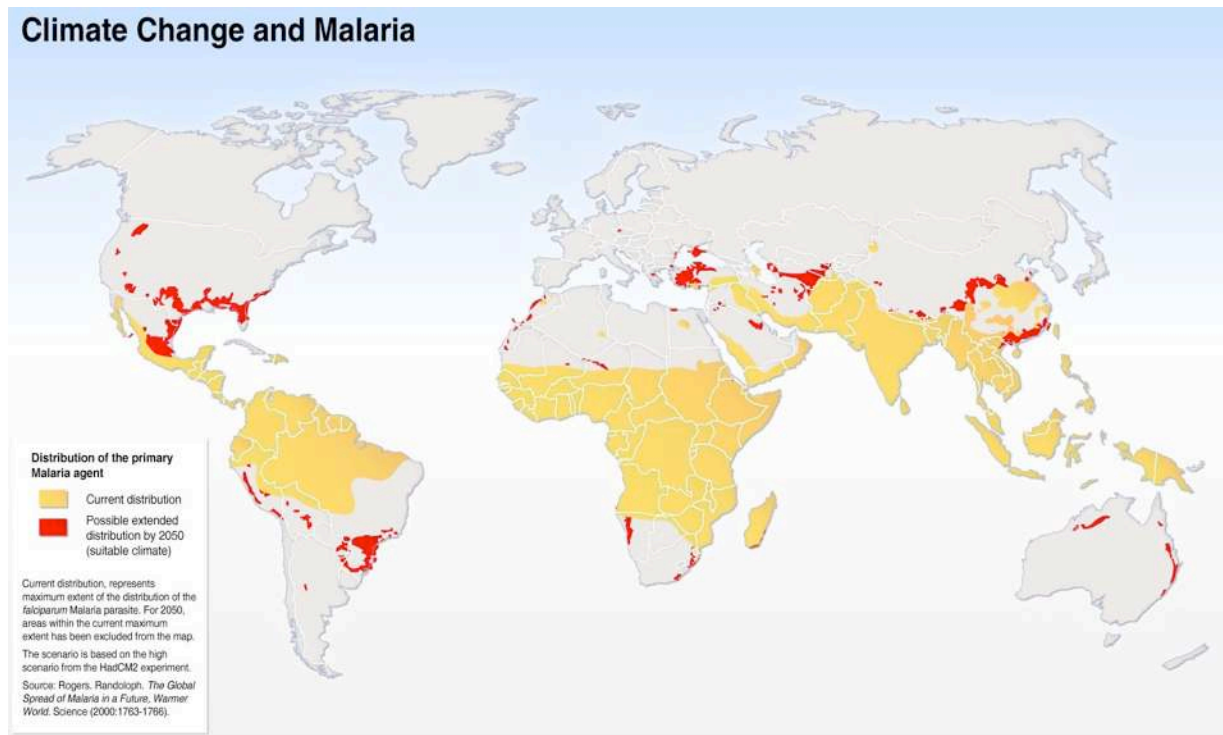


Figure 1.1. Forecasted impact of climate change on the worldwide spread of malaria in 2050. Climate change will allow malaria to spread into new areas. This map shows the new areas where the Malaria parasite *Plasmodium falciparum*, will likely be able to spread by 2050 based on the Hadley Centre model's high scenario. Areas shown in yellow indicate the current distribution of malaria. Areas shown in red indicate areas where climate will be suitable for malaria by 2050. Other areas may become free of malaria as climate changes.

Courtesy of Hugo Ahlenius, UNEP/GRID-Arenda

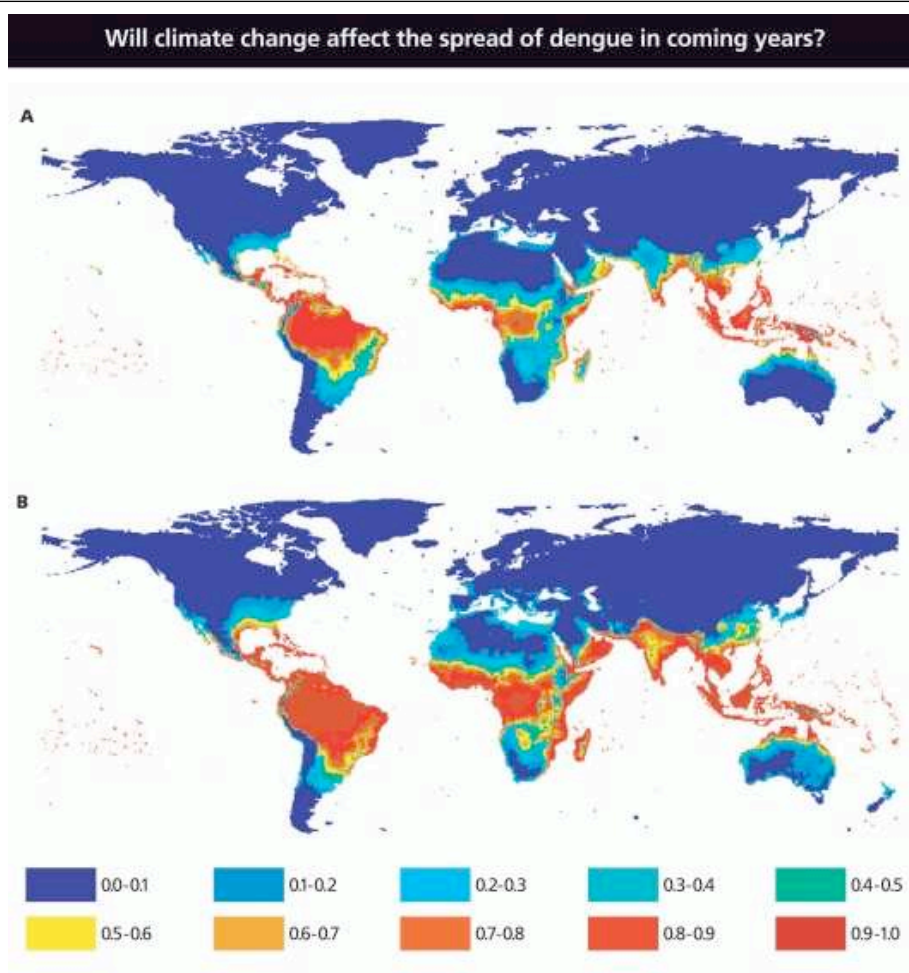


Figure 1.2. Estimated baseline population at risk for dengue infection in 1990 (A) and in 2085 (B) based on modelling using climate data for 1961–1990 and projections for humidity change a function of climate change—for 2080–2100. Ranges above indicate percentage of the population at risk: 0–10%, 10–20%, etc. However, many scientists do not agree that climate change will appreciably alter the risk of dengue. Source: Hales S, et al. 2002.

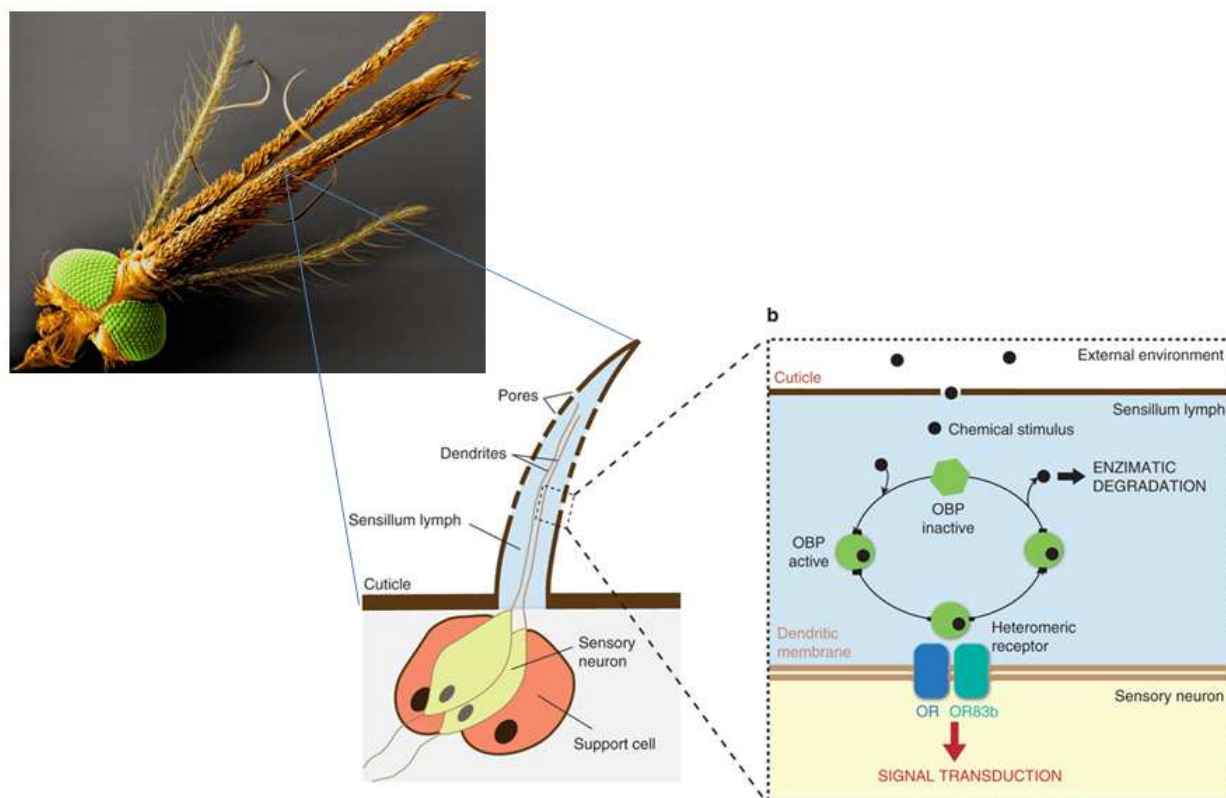


Figure 1.4. Details of the sensory organs and tissues that are components of the olfactory system in mosquitoes. (a) A colored scanning electron microscope image of a female malaria mosquito's head shows its impressive array of olfactory sensors. The two feathery outer appendages are the antennae. The proboscis is in the middle, flanked by the maxillary palps that specialize in detecting odors coming from human hosts. (Credit: Zwiebel Laboratory) (b) Schematic representation of the general structure of an insect olfactory hair. Gustatory sensilla have a similar structure, with only a single pore at the top of the sensory hair. The first molecular steps (perireceptor events) of the insect chemosensory signalling transduction pathway. This figure depicts a general, simplified functional scheme; alternative schemes for OBP activity have been proposed (see below). Source : Sánchez-Gracia.A et al. (2009).

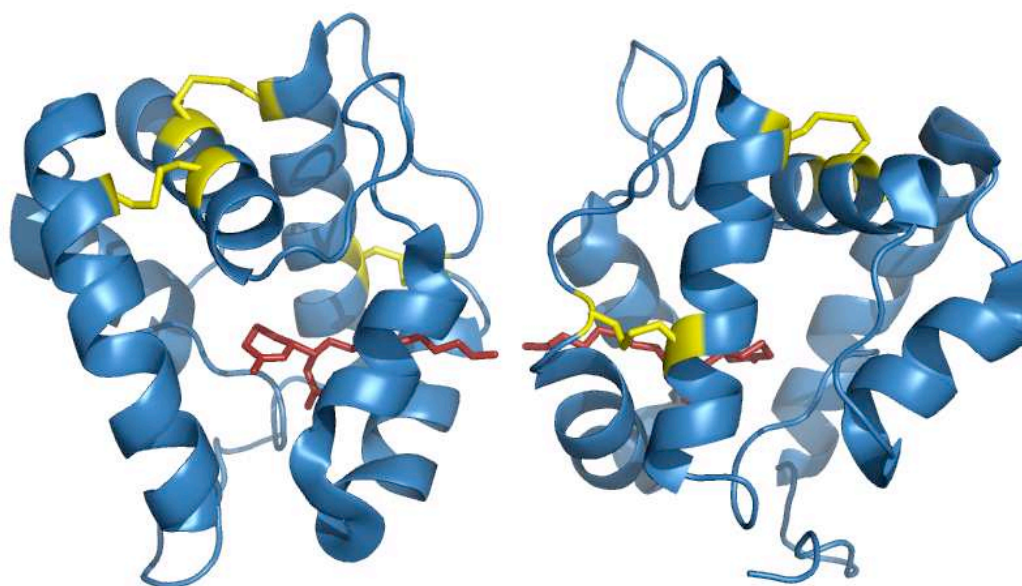


Figure 1.5. Structure of CquiOBP1 dimer from *Culex quinquefasciatus* bound to ‘3OG: (1S)-1-[(2R)-6-oxotetrahydro-2H-pyran-2-yl]undecyl acetate’ (colored in red). The Cysteines involved in Disulphide bond formation are indicated in yellow. PDB ID: 3OGN.

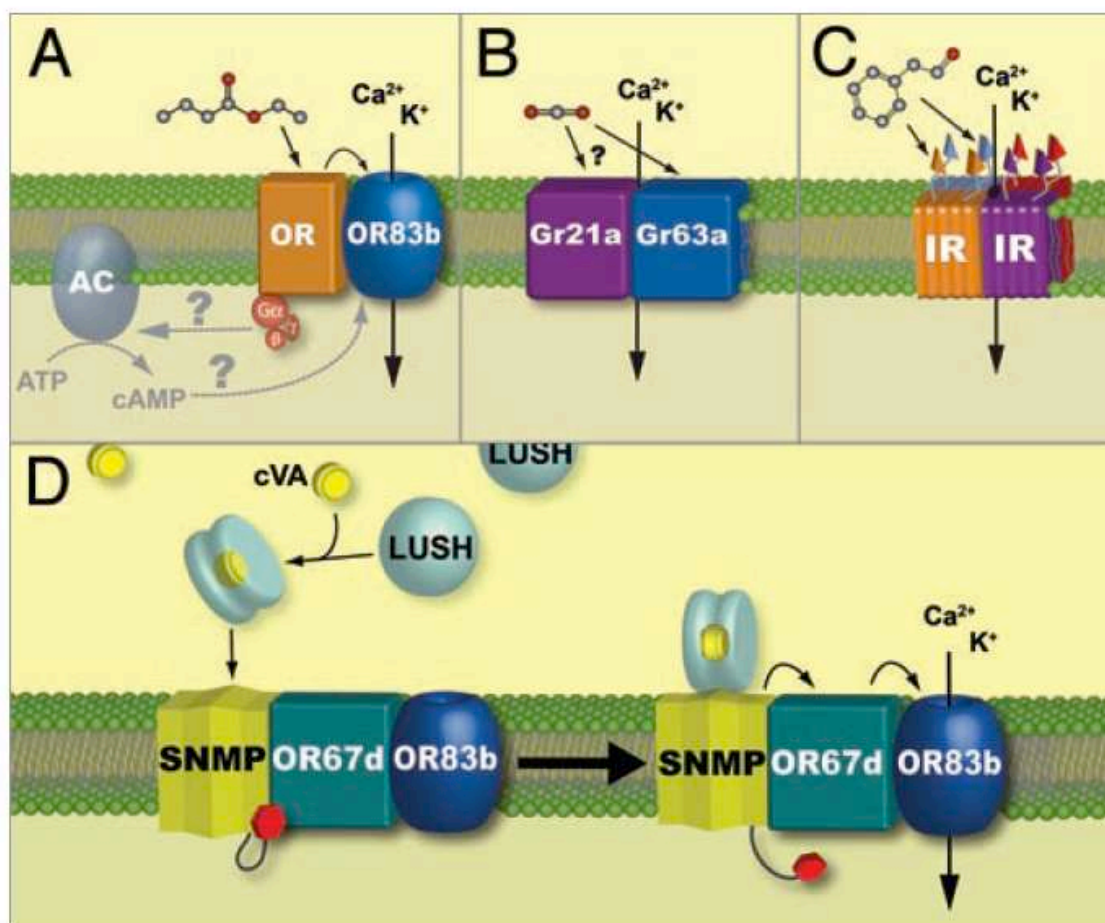


Figure 1.6. Mechanisms by which odorants are detected in *Drosophila*. (A) A ‘tuning’ OR (that confers odorant specificity to the neuron) forms a heterodimer with the ion channel Or83b. Food odorants (shown as the ball and stick structure) interact with the ‘tuning’ OR, which in turn activates the Or83b ion channel. Or83b conducts potassium and calcium ions into the olfactory neurons, resulting in depolarization and initiation of action potentials. Thus, the OR/Or83b heterodimers are ligandgated ion channels. Odorant binding to the OR may also trigger activation of Or83b via a second messenger, wherein a G protein (in red) stimulates adenylyl cyclase (AC) to produce cAMP, which in turn activates Or83b. (B) CO₂ detection is mediated by a heterodimer of gustatory receptors (GRs) Gr21a and Gr63a; taste receptors expressed in ac1c basiconic OSNs of the antenna. (C) Variant ionotropic glutamate receptors (IRs), mediate odorant detection in coeloconic sensilla. The extracellular ligand-binding domains of IRs (tethered triangles) likely recognize odorants and activate the channels, which are likely to be heteromultimers. (D) Detection of cVA pheromone (yellow discs) is mediated by the extracellular receptor LUSH (blue discs). cVA binds LUSH, inducing a conformational shift in LUSH (shown as the LUSH dimer with cVA bound). Activated LUSH binds the neuronal receptor complex consisting of SNMP/Or67d/Or83b, thereby activating the ion channel Or83b. SNMP may function as an inhibitory subunit of Or67d/Or83b in the absence of activated LUSH. Upon LUSH/cVA binding, SNMP releases inhibition on the receptor complex and also aids in further activation of Or67d/Or83b Source : Ronderis and Smith, 2009.

2

Genomic characterization of odorant binding proteins in three mosquito genomes

2.1. Introduction

The spread of infectious diseases among humans is mediated primarily by the world's most dangerous animal, the mosquitoes among which the anthrophilic mosquitoes such as *Anopheles gambiae*, *Anopheles funestus*, *Aedes albopictus*, *Aedes aegypti*, *Culex quinquefasciatus* are the most effective transmitters of viruses and parasites. They are responsible for the spread of a number of life threatening diseases such as malaria, dengue, and West Nile encephalitis and recently *Chikungunya*. According to the World Health Organization, global climate change is expanding mosquitoes range, heightening the risk of disease for millions of additional people. Primary prevention is one of the most important aspects to subside the spread of diseases either by controlling the population of these vectors or by preventing the interaction between the vector and the host.

Understanding the molecular mechanism for human host recognition mediated by olfaction would help in identifying new strategies for the prevention of the primary contact. Volatile products secreted by the human host in the process of metabolism are responsible for the attraction of these vectors to the host. The ability of recognizing and discriminating thousands of odorant molecules in insects as in mammals relies on specialized chemosensitive neural cells expressing olfactory receptor proteins (ORs) which reside within segregated compartments called sensilla. Each sensillum is a hair like structure bathed in the sensillum lymph which contains a number of secreted proteins (McKenna et al. 1994; Pikielny et al. 1994; Wang et al. 1999). The odorant binding proteins (OBP) are found to be important water-soluble components of this sensillum lymph. It was first identified in the moth as pheromone binding proteins or PBP (Vogt and

Riddiford 1981). These globular proteins are believed to bind different odorant molecules (Plettner et al. 2000) owing to their high divergence within the family and transport them to their respective olfactory receptors triggering the mechanism of olfaction (Pelosi and Maida 1995).

The arthropod odorant binding proteins (OBPs) form a large specific multi-gene family. They are 10-30kDa globular and water-soluble proteins that are characterized by a specific 6- α helices domain comprising of six highly conserved cysteines that have distinct disulphide connectivities and which are now considered the hallmark of this protein family (Calvo et al. 2002; Valenzuela et al. 2002; Calvo et al. 2006). OBPs have been identified in a number of insect species including four dipterian species *Drosophila melanogaster* (Galindo and Smith 2001) and (Graham and Davies 2002; Hekmat-Safe et al. 2002; Valenzuela et al. 2002; Zhou et al. 2004; Vieira et al. 2007; Vieira and Rozas 2011), *Anopheles gambiae* (Vogt et al. 2002; Xu et al. 2003; Zhou et al. 2004; Vieira and Rozas 2011), *Aedes aegypti* (Zhou et al. 2008) and *Culex quinquefasciatus* (Pelletier et al. 2010). These proteins are very divergent in terms of the sequences within the family and sequence identities of the family members among the different species could drop as low as 8% (Vieira and Rozas 2011). In the *Drosophila*, a subgroup of (i) odorant binding proteins lacking 2 of the 6 conserved cysteines, called *Minus C* OBPs and (ii) OBPs carrying additional conserved cysteines called *Plus C* OBPs have been identified (Hekmat-Safe et al. 2002). The *Minus C* OBPs typically lack the second and fifth Cys residues. However this definition appears to be somewhat ambiguous, since there are three *Drosophila* OBPs among this cluster which contain all the six hallmark cysteines (Pelosi and Maida 1995). *Minus C* OBPs have never been described so far in mosquito genomes.

Three subfamilies of OBP genes have been identified so far among mosquitoes : (i) the *Classic* OBPs carry the six conserved cysteines characteristic motif of the odorant binding protein family; (ii) the *PlusC* OBPs have the same conserved cysteines and disulphide connectivity but which contain six additional cysteines with novel disulphide connectivities; (iii) the *Atypical* OBPs are among the longer known OBPs and have initially been described as containing a single *Classic* OBP domain in its N-term extended by a less characterized C-term extension. Very recently, it was shown that *Atypical* OBPs are composed of two domains that are in fact homologous to the *Classic* OBP domain and were hence considered as “Dimer” OBPs (Manoharan et al. unpublished; Vieira and Rozas 2011)

In *A. gambiae* and *A. aegypti*, OBPs from the three different subfamilies have been reported to date while in *Culex quinquefasciatus*, only the *Classic* members of this family have been

reported so far (Pelletier and Leal 2009). *Atypical* and *Plus C* OBPs have not been reported yet in this genome.

An additional multi-gene family, known as D7 salivary proteins are known to be distantly related to the arthropod odorant binding protein superfamily (Calvo et al. 2002; Calvo et al. 2006; Calvo et al. 2009). There are two types of D7 salivary proteins in the mosquito genome, the short and the long forms which contain a single and two OBP-like domains respectively (Valenzuela et al. 2002; Kalume et al. 2005; Choumet et al. 2007). The available structures of the D7 proteins indicate that the domains adopt a similar fold to the OBP domains but decorated with additional structural features and a seventh helix. In the two-domain D7 protein, the C-term OBP-like domain have been shown to bind to biogenic amines in *A.gambiae* and *Aedes aegypti* (Mans et al. 2008; Calvo et al. 2009) while N-terminal domain in *Aedes aegypti* was shown to have a specific bioactive lipids binding activity (Calvo et al. 2009). These members serve as important representatives for the construction of phylogenetic trees serving as outgroups for the OBP gene family in the current analysis.

The identification of *Plus C* and *Atypical* members of the *Culex quinquefasciatus* genome and an expansion of their *Classic* OBP members in addition to those identified by (Pelletier and Leal 2009) is reported in this analysis. The existing dataset of *A.gambiae* and *A. aegypti* (Vogt et al. 2002; Xu et al. 2003; Zhou et al. 2004; Zhou et al. 2008; Vieira and Rozas 2011) is enriched with new entries identified by the methods used in the analysis. The analysis also reveals the identification of *MinusC* OBPs closely related to the *Drosophila MinusC* OBPs (Hekmat-Safe et al. 2002) and additional true *MinusC* proteins which lack C2 and C5 cysteines and which are homologous to *Bombyx mori MinusC* proteins (Zhou et al. 2009). The characteristics of odorant binding protein subfamilies in each of the genome using structure based alignments and phylogeny with the help of the distantly related D7-related family of proteins as an out group was analyzed. An analysis has also been extended on the comparison of the different classes of the OBPs among the mosquito genomes along with the *Drosophila* odorant binding proteins (Hekmat-Safe et al. 2002) in the case of *Classic* OBPs.

The *Atypical* odorant binding proteins form a unique family of odorant binding protein gene family first being identified in the *Anopheles gambiae* genome with sixteen members (Xu et al. 2003; Zhou et al. 2008). Subsequently sixteen members of *Atypical* odorant binding proteins were identified in the *Aedes aegypti* genome (Zhou et al. 2008). However sensitive sequence search methods used here indicated the presence of an additional 31 *Atypical* members in the *Aedes aegypti* genome. The presence of 26 *Atypical* members in the *Culex quinquefasciatus* genome is

also being reported for the first time. Analysis on the current enriched dataset of the *Atypical* family in the mosquito genome has helped in the classification of these members into different subtypes based on comparative genome analysis. The different subtypes have been named *matype1-4*. Unique cysteine conservation patterns among these members have been found that were not observed earlier. The extended C-terminal region was found to have additional 6 conserved cysteines (C1'-C6') in the *matype 1, 3, 4* making a total of 12 conserved cysteines. The cysteine spacing of these additional cysteines was found to be similar to the spacing observed in the first 6 conserved cysteines (C1-C6) that corresponds to the *Classic* OBP fold. The current analysis confirms that all *Atypical* members are indeed two-domain OBPs and closest homologues to each domains based on Psi-Blast search are indicated. In addition the *matype 2* members with an exception of the members from *Anopheles* were found to lack C2, C5, C1', C2', C3', C5' making a total of only 6 conserved cysteines suggesting that they are truly *Atypical* members.

2.2. Materials and methods

2.2.1. Sequence searches

The predicted protein sequence database of the three mosquito genomes *A.gambiae* (<http://www.vectorbase.org>, *Anopheles gambiae* annotation, AgamP3.4), *A.aegypti* (<http://www.vectorbase.org>, *Aedes aegypti* annotation, AaegL1.1) and *C.quinquefasciatus* (<http://www.vectorbase.org>, *Culex quinquefasciatus*, CpipJ1.2) were downloaded from the Vectorbase (Lawson et al. 2009). The putative odorant binding proteins in the three mosquito species were identified using 10 *Drosophila* query sequences which belong to three different subfamilies Classic/General OBPs, *Plus C* and *MinusC* OBPs using a PSI-BLAST (Altschul et al. 1997) run of 10 query sequences with an E-value cutoff of 3e-10 (Vieira et al. 2007) and a alignment length cutoff of 75% with respect to the query sequence. At this level, all of the previously identified members in the three genomes were identified with identification of a few additional members. A second run of Psiblast was initiated with the hits from the previous runs. Using this protocol it was possible to not only pick up all the members of OBPs reported so far (Vogt et al. 2002; Xu et al. 2003; Zhou et al. 2004; Zhou et al. 2008; Pelletier and Leal 2009; Vieira and Rozas 2011) but also a remarkable number of additional members. The additional sequences were checked for the presence of PBP/GOBP domain in the case of general odorant binding proteins and alignment of the new sequences with their subfamily members in case of *Atypical* and *Plus C* proteins. The D7 proteins identified using this method were also retained for further analysis and used as an outgroup in the construction

of phylogenetic trees. The orthologous sequences were identified based on the reciprocal best hit approach using BLAST (Moreno-Hagelsieb and Latimer 2008). The newly added sequences are named according to the naming conventions used in the earlier reports (Vogt et al. 2002; Xu et al. 2003; Zhou et al. 2004; Zhou et al. 2008; Pelletier and Leal 2009)

2.2.2. Multiple sequence alignment

The multiple sequence alignment forms the basis for any analysis of a family of proteins and it is highly necessary to obtain an accurate alignment. The error rate in the alignment increases with the increase in divergence of the proteins. Structure-based alignments in turn are considered to be the most accurate forms of alignments and hence, in this study, the structure alignment was used in constructing the alignments. The structure alignment was constructed using 10 odorant binding proteins in the OBP gene family using COMPARER (Sali and Blundell 1990) (data not shown). However the use of the structure alignment as profiles was restricted to seven members in the case of OBPs and 2 members for the D7 family due to the limited number of structural data. The OBPs and the D7 sequences were aligned to their respective structure alignments as profiles and a combined alignment of the two family of proteins was constructed using the profile-profile alignment option using Clustal X (Thompson et al. 1994; Thompson et al. 1997; Jeanmougin et al. 1998). The alignments were truncated based on the structure alignment on the N-terminal end however the C-terminal ends were retained due to the presence of an extended C-terminal in the case of *Atypical* subfamily members of the OBP family. This method was applied for aligning the sequences in all the three different genomes. Alignments for the different subclasses were constructed with sequences from all the three mosquito genomes and in the case of *Classic* subfamily, along with *Drosophila* sequences. The alignment of the *Atypical* and *Plus C* subclasses of OBPs were however not based on the structure alignment.

2.2.3. Phylogenetic analysis

The phylogenetic trees were inferred using the Neighbor-Joining method (Saitou and Nei 1987) in MEGA 4.0 (Tamura et al. 2007). The percentage of replicate trees in which the associated sequences cluster together in the bootstrap test (1000 replicates) are shown next to the branches (Felsenstein 1985) and branches with less than 50% bootstrap cutoff were collapsed. The evolutionary distances were computed using the Poisson correction method (Zukerkandl and Pauling, 1965) and are in the units of number of amino acid substitutions per site. All positions

containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons (pairwise deletion option). The trees were rooted at the branches of the D7 family of proteins which was considered as an outgroup. The trees of the different subclasses used for the comparative analysis of the different genomes were analyzed as unrooted trees.

2.2.4. Chromosomal mapping

The figures of the chromosomal mapping were drawn to scale using Adobe illustrator. The genes were mapped to their respective location on the chromosome or supercontigs. The chromosome of *Anopheles gambiae* is used as reference and are represented as a yellow bar and the contigs of *Aedes* and *Culex* are represented in purple and green respectively (supplementary Figures 1a-e). The direct three-way (1:1:1) orthology relationships among the three genomes are represented as green lines. The two-way (1:1) orthology relationships between two species are represented as black lines and the inparalogy relationships are represented as red lines.

2.3. Results

2.3.1. Naming of OBP genes from mosquitoes needs to be clarified

In this study, the same naming convention that was adopted in previously published works was used and it also relied on the OBP names used in VectorBase which by itself is based on published data. In particular, the most recent OBP genes (AgamOBP58 to AgamOBP68) from the *Anopheles gambiae* genome were named after Zhou et al (2008), even though these authors did not provide any gene accession numbers or *id* nor their chromosomal localization. The tentative identities of these *An. gambiae* genes have been identified by manual inspection of their corresponding sequences provided in the supplementary material of their publication. Besides, it is noteworthy that, in a few cases, the names of OBP members from *Anopheles gambiae* used in this study did not coincide with the gene names published very recently by Vieira and Rozas (2011). A comprehensive comparative list of the gene names adopted in the respective works is provided in supplementary Table 1a and 1b.

A few genes in the *Anopheles gambiae* genome have been removed after identifying and removing a few duplicate entries (see below and supplementary Tables 1a-d). In *Aedes aegypti* a naming ambiguities for few genes were also identified in the literature and are explicated below.

The *Classic* OBPs AgamOBP6 and AgamOBP18 (Xu et al. 2003) are identical genes with identical location (vectorbase ID: AGAP003530). Hence the name AgamOBP6 was retained. This

naming ambiguity is indicated in supplementary Tables 1a and 1b. Similarly, the previously reported AgamOBP15 and AgamOBP16 (vectorbase ID: AGAP003307) from (Xu et al. 2003) were identified to be identical in sequence and location. Therefore the gene was designated as AgamOBP15 like in Vieira and Rozas (2011). The gene (vectorbase ID: AGAP003309) was previously reported as AgamOBP1 and AgamOBP17. In the VectorBase database the gene is referred as AgamOBP17 and AgamOBP1 is absent while in Vieira and Rozas (2011), it is referred as AgamOBP1. Importantly, it should be noted that PDB structures of the gene product refers to the name AgamOBP1 and in related publications. The closest homologues of AgamOBP1 are also referred as AaegOBP1 and CquiOBP1 in the literature. This gene is finally renamed as AgamOBP1 in this study. The gene AGAP002905 is referred as AgamOBP14 and AGAP002189 as AgamOBP13. However, in Vectorbase, these genes names are inverted. This is indicated in supplementary Tables 1a and 1b. As discussed above, in their paper Zhou et al (2008) claimed to have identified new sets of OBP genes that were named AgamOBP58-68. Even though the sequences were provided in their supplementary material in the form of sequence alignments, neither gene ID or accession number data were provided in their paper nor their corresponding genomic localization as they did for the *Aedes aegypti* genes in the same paper. This rendered difficulty in the identification of these genes. The identification of these genes have been attempted by manual inspection of their corresponding sequences extracted from their supplementary material. Nevertheless, a few sequences reported by Zhou et al (2008) have been removed from this study since a few members were found to be duplicates of already reported entries : (i) their described AgamOBP59, AgamOBP60 and AgamOBP61 were removed from the final list (supplementary Tables 1a and 1b) as they are identical to the genes AgamOBP49, AgamOBP51 and AgamOBP53 respectively that were already reported by Xu et al. (2003); (ii) their AgamOBP65 was removed as it was identical to AgamOBP7 (Xu et al. 2003). As a result, the genes AgamOBP62-AgamOBP64 from their analysis have been renamed as AgamOBP59-61 and AgamOBP66-68 are called AgamOBP62-64 respectively. These sets of genes have gene names that do not coincide between the current study and the work of Vieira and Rozas (2011). A detailed comparison of the naming used by Vieira and Rozas (2011) and the naming used in this study for the *Anopheles* OBP gene products is provided in the supplementary Table 1b.

In the *Aedes aegypti* genome, all the OBP in this study were named after Zhou et al (2008) except for their genes AaegOBP39, AaegOBP27 and AaegOBP56 which were in fact named as AaegOBP1, AaegOBP2 and AaegOBP3 respectively accordingly to the names given by Ishida et al. (2004) who reported them for the first time. This was further corrected in a *corrigendum* of Zhou et

al (2008) paper where this error was rectified. The gene names and corresponding gene ID/accessions from *Aedes aegypti* are listed in supplementary Table 1c.

As for the *Culex quinquefasciatus* OBP genes, the naming convention that is adopted here is from Pelletier et al. (2009). The detailed list of the OBP genes from *Culex quinquefasciatus* are provided in supplementary Table 1d.

2.3.2. Extension of odorant binding proteins family in all 3 mosquito genomes

In the already published works, 64 odorant binding proteins from *A. gambiae* (Vogt et al. 2002; Xu et al. 2003; Zhou et al. 2004; Zhou et al. 2008; Vieira and Rozas 2011), 66 from *Aedes aegypti* (Zhou et al. 2008) belonging to the three known OBP gene subfamilies and 53 *Classic* OBPs from *Culex quinquefasciatus* (Pelletier and Leal 2009) have previously been identified. Only very recently, Vieira and Rozas (2011) added 5 new putative genes to the *Anopheles gambiae* OBP gene repertoire.

In this study, new OBP sequences from the three mosquito genomes were identified using the sequence search approach described in materials and methods. In total, the identification of four new OBPs in *A.gambiae* and 47 new OBPs in *A. aegypti* (Table 1) with respect to Zhou et al (2008) are reported. 61 new OBPs in *C. quinquefasciatus* with respect to Pelletier et al (2009) (Table 1) are also reported. These new entries are detailed below and in supplementary Tables 1a-d.

Classification of these new OBPs into the different subfamilies was performed after aligning them to previously identified members (see complete details in subsequent subsection). For *A. aegypti*, 6 new *Classic* OBPs have been identified among which AaegOBP78 does not have the 2nd and 5th cysteine but is predicted to have the GOBP domain. Similar proteins missing these cysteines were also identified in the *C. quinquefasciatus* genome (see below). Additionally 10 new members in the *Plus C* OBP group in *Aedes* genome (AaegOBP67-75 and AaegOBP82) have been identified. As for the OBPs that fall into the *Atypical* class, 31 new members (AaegOBP84 to AaegOBP114) are identified which interestingly show high sequence similarities with the new *Atypical* members from the *C. quinquefasciatus* genome that have been reported for the first time in this work (see below).

In the case of *A. gambiae* the identification of four new OBP sequences (AgamOBP65, AgamOBP66, AgamOBP67 and AgamOBP68) which Vieira and Rozas (2011) also recently identified as AgamOBP66, AgamOBP62, AgamOBP67 and AgamOBP68 respectively are confirmed in this study (supplementary Tables 1a and 1b). AgamOBP65 and AgamOBP66 are novel

PlusC OBP that have not been described before in *Anopheles* genome while AgamOBP67 and AgamOBP68 which are also *PlusC* OBPs are located on the “unknown” chromosome of the *Anopheles* genome and have 100% sequence identity to AgamOBP61 and AgamOBP66 respectively.

For *Ae. aegypti*, 6 new *Classic* OBPs (AaegOBP76-81 and AaegOBP83) have been annotated among which AaegOBP78 does not have the 2nd and 5th cysteine but is predicted to have the GOBP domain. Similar proteins were also identified in the *C. quinquefasciatus* genome (see below). In addition, 10 new members were added to the *Plus C* OBP group (AaegOBP67-75 and AaegOBP82). In the *Atypical* class of the *Ae. aegypti* genome, 31 new members have been identified (AaegOBP84 to AaegOBP114), which show high sequence similarities with members found in the *C. quinquefasciatus* genome.

In *C. quinquefasciatus* genome, 53 members were reported in the *Classic* group by Pelletier *et al.* (2009). Here an extension of the *Classic* OBPs with 21 additional members (CquiOBP54-CquiOBP74) and also the identification of 26 *Atypical* sequences (CquiOBP75-CquiOBP100) and 12 *PlusC* odorant binding proteins (CquiOBP101-CquiOBP112) (Table 1 and supplementary Table 1d) is reported. The members of *Atypical* and *PlusC* proteins have never been described previously in the *C. quinquefasciatus* genome.

Among the newly added sequences in the *Classic* OBP class, the 2nd and 5th cysteines were not conserved for 15 sequences (CquiOBP59-CquiOBP62, CquiOBP64-CquiOBP74) but they were found to carry the GOBP/PhBP domains with significant values in the CDD (Conserved Domain Database) search and were retained as putative OBP members. The VectorBase entries for the CpijOBP45 and CpijOBP47-50 reported by Pellitier *et al.* (2009) were not found in the genomic data available in version 3.4 of VectorBase and have not been used for the sequence alignment and phylogenetic analysis in this study.

2.3.3. Alignment of OBP proteins and description of their key sequence features

Owing to the low sequence identity and length variations observed between the members of the OBP family, a structure-based alignment was used as a guide to align them (see materials and methods). This approach is being used for the first time for the analysis of OBP multi-gene family among insects. It highly improved the quality of alignment compared to regular multiple sequence alignments namely for (i) the precise classification of the new OBPs into the three different

subfamilies and (ii) the identification of residues in structurally conserved positions that would have been missed otherwise (supplementary Figures 2a-2c).

The conservation pattern of cysteines across the different classes were clearly highlighted in these structure-based alignments but could not be obtained with the ordinary sequence alignments methods. The cysteine positions in this chapter are referred by numbering them C1 to C6 with respect to the order of their positions in the *Classic* OBP proteins. A detailed schematic representation featuring the cysteine spacings, conservation together with their predicted disulphide patterns are given in Figure 1. Overall, the six cysteine residues involved in disulphide bond formation which are considered as the hallmark of this protein family (Calvo et al. 2002; Valenzuela et al. 2002; Calvo et al. 2006) are very well conserved across the *Classic*, *Plus C* and *Atypical* subclasses.

Interestingly, the *MinusC* subtype that falls within the *Classic* OBPs and that are reported as “*Bombyx mori (minus C)*” subclass is being described for the first time in the mosquitoes genomes *A.aegypti* (AaegOBP78) and *C.quinquefasciatus* (CquiOBP59-CquiOBP62, CquiOBP64-CquiOBP74). This subtype was not found in *A. gambiae* and seems to be restricted to the *Culiniidae* species.

The very recent and preliminary observation by Vieira and Rozas (2011) that *Atypical* OBPs in mosquitoes should be considered are “dimer” OBPs because they contain a second GOBP domain with the six hallmark cysteines is confirmed in this analysis. Indeed, the close analysis of the extended C-terminal end of *Atypical* members highlighted the presence of 6 additional cysteines conserved within this subtype which hold a cysteine spacing pattern very similar to the conserved cysteines (C1-C6) at their N-terminal end. Hence it is proposed that these cysteines are annotated as C1'-C6' and it is noteworthy that within the *Atypical* subfamily, a distinctive subtype called *matype2* (see below and figure 1) showed the presence of only 6 cysteines (C1, C3, C4, C5, C14', C6') when compared to the other subtypes which carry the 12 cysteines.

As expected and as already reported previously, sequence divergence is high among OBP family members. The average sequence identity between OBP genes in *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* are 12.5%, 12.8% and 13.1% respectively. Genes from these three genomes that are from the *Classic* subfamily share an average sequence identity of 15.5% while the *PlusC* and *Atypical* genes share respectively 17.3% and 22.1% average sequence identity. The corresponding values within each phylogenetic subgroup are further detailed in supplementary Figure 1. Some subgroups are characterised by a very low average sequence identity like the *Bombyx mori minusC* (21.5%), the *mclassic9* (23.3%) or the *mplus7c* (24.3%) while other

subgroups share significantly higher sequence identities like OS-E/OS-F (55.2%), Pbprp4 (60.2%) or mclassic4 (77.3%).

It is difficult to distinguish conserved sequence features other than the cysteine residues on a global basis, *i.e* if one considers an alignment with all the OBPs from one genome. Nevertheless, close examination of the alignments for the different subgroups in the subfamilies (supplementary Figures 2a-2c) indicates that the phylogenetic clusters established in this study tend to have specific sequence patterns. In Figure 2a-c are illustrated the conserved residues within each subfamily in the form of sequence logos. Interestingly, for the *Atypical* OBPs, specific conserved residue motifs can be clearly identified while this is not evident for the *Classic* and *PlusC* OBPs due to high sequence divergence.

2.3.4. Analysis of OBP genes orthology across the 3 genomes and their corresponding distribution

The orthology and the gene distribution of the OBPs in the three genomes was investigated. Assembled genome is only available for *Anopheles gambiae* at the date of this work and in version 3.4 of VectorBase. The chromosomal mapping for every of the OBPs genes in *Anopheles* is hence known with precision. Their chromosomal distribution in the *Anopheles* genome are centrally featured in supplementary Figure 1 and further referenced in supplementary Tables 1a, 1c and 1d. Though the syntenic relationship between the chromosome arms in *Anopheles gambiae* and their corresponding orthologous chromosome arms in *Culex* and *Aedes* was established by Arunsberger *et al*, (2010) with the help of genetic markers (supplementary Table 2), the genomic data of these two Culinidae species are only available in the form of supercontigs fragments (Nene et al. 2007; Arensburger et al. 2010) and have yet to be assembled. In these two genomes, a few supercontigs (about 10%) harbor markers that allow their chromosomal localization (Arensburger et al. 2010). Very few of these anchor supercontigs hosted OBP genes. Most supercontigs containing OBP genes did not harbor any genomic markers, hence cannot be assigned to a chromosome in *Aedes* and *Culex*. However, in many cases, direct orthologues in the *Anopheles* genome could be identified (supplementary Figures 1a-e and supplementary Tables 1a, 1c and 1d). OBP orthologues have been identified using the reciprocal blast hit approach (Moreno-Hagelsieb and Latimer 2008) which is widely used in the detection of orthologues. As illustrated in Figure 3 and supplementary Figures 1a-e, three way orthology (1:1:1) between OBP genes in the three genomes were identified in 31 cases while two way orthology (1:1) between OBP genes from only two genomes were identified in 5 cases between *Anopheles* & *Culex*, 6 between *Anopheles* & *Aedes* and 19 between *Aedes* and

Culex (Figure 3) thus confirming the genetic proximity between the *Aedes aegypti* and *Culex quinquefasciatus* species.

Interestingly, the overwhelming majority of the OBP genes are organized in gene clusters in the 3 genomes (supplementary Figures 1a-e). The clusters are mainly composed of gene duplicates. The genes in these clusters hence share high sequence identity (data not shown) and are thereby phylogenetically very close (see below) as it is confirmed by inParalogy data from the inParanoid database (O'Brien et al. 2005). The extension of OBP gene repertoire in *Aedes aegypti* and *Culex quinquefasciatus* with respect to *Anopheles gambiae* was mainly driven by these gene duplications events which are more numerous in these two Culinidae species. There are a total of 12 OBP genes clusters in *Aedes aegypti* and 13 clusters in *Culex quinquefasciatus* genomes when compared to 6 clusters in *Anopheles gambiae*. The largest gene clusters are found in *Aedes* and *Culex* and a few clusters contain as much as 12 genes. It is observed that 21 out of the 26 newly identified *Atypical* OBPs genes from *Culex quinquefasciatus* are in fact distributed into 3 main gene clusters (supplementary Figures 1a-e). Similarly, 10 out of the 12 newly identified *PlusC* proteins are distributed into 3 gene clusters.

2.3.5. Phylogenetic analysis of the odorant binding proteins

Rooted phylogenetic trees were constructed for each individual mosquito genome to study the divergence of OBPs and to analyze the clustering of the newly identified members into their respective classes. The alignments used for the construction of phylogenetic trees were obtained by aligning the sequences to a high quality structure based alignment of the odorant binding protein structures available in the PDB (August 2009). The structure alignment was constructed using COMPARE (Sali and Blundell 1990) and included 7 *Classic* OBPs and two D7 proteins (data not shown). The D7 sequences and *Classic* OBP sequences were aligned separately with their respective structural alignments as profiles generating a D7-related profile and an OBP profile. These two profiles were subsequently aligned using a profile-profile alignment method. Phylogenetic trees were then constructed using the neighbour joining method (Saitou and Nei 1987) and MEGA 4.0 software (Tamura et al. 2007) with thousand bootstrap replicates and D7 proteins were used as an outgroup to root the trees.

The use of structure alignment highly increased the quality of alignment compared to regular multiple sequence alignments namely for the identification of residues in structurally conserved positions that would have been missed otherwise. The derived conservation patterns had

a huge impact on the branching of the phylogenetic trees in particular the conservation pattern of cysteines across the different classes was clearly highlighted in these alignments but could not be obtained with the ordinary alignments. These alignments and the use of D7-proteins as an outgroup also impacted upon the robustness of the phylogenetic trees by increasing the bootstrap values of the different subclasses of the OBPs within the genome, which could not be obtained otherwise.

Overall, for *A.gambiae* and *A.aegypti*, the results highly coincided with the previous reported phylogeny of the OBPs however with additional members and stronger bootstrap support. The branching patterns also confirm that the *Atypical* sequences are more closely related to the *Classic* odorant binding proteins than the *Plus C* odorant binding proteins. The phylogeny of the OBP gene family from the *C.quinquefasciatus* genome with the newly identified *Atypical* and *Plus C* subclasses is also reported.

2.3.6. Comparative analysis of the *Classic* and *Plus C* subfamilies of OBPs

Comparative analysis of proteins among different species helps in better understanding their evolution and also in identifying the function of the proteins. It also helps in identifying sequences specific to a particular genome corresponding to a specific genome. The comparative analysis of the different subfamilies of the OBPs in the mosquito genome helped in observing the clustering patterns within each subfamily of the OBP members. The analysis was done based on the sequence alignment and phylogenetic trees constructed using sequences from individual subfamilies from all the three mosquito genomes used in this analysis and the *Drosophila* OBPs (Hekmat-Scafe et al. 2002) in the case of the *Classic* members. A consensus tree was constructed for 178 sequences using the neighbour joining method (Saitou and Nei 1987) with all the *Classic* odorant binding proteins from the three mosquito genomes and the *Drosophila melanogaster* with 1000 bootstrap replicates. The clustering of the various *Classic* OBPs into significant clusters revealed the possibility of 19 different subtypes. Sequences were aligned based on the structure alignment of the OBPs mentioned earlier. Few members of the mosquito genomes clustered with *Drosophila* OBPs (Hekmat-Scafe et al. 2002) and were named after the *Drosophila* OBPs in the respective clusters. Among these OSE-OSF, Pbprp1, Lush, OBP19a and Pbprp4 have already been described in (Pelletier and Leal 2009). However one member from *Culex quinquefasciatus* in each of the two subtypes OSE-OSF (CquiOBP58) and OBP19a (CquiOBP57) have been annotated. The huge expansion of sequences (CquiOBP25-CquiOBP42) observed by (Pelletier and Leal 2009) were found to be homologous to AegOBP57 and AgamOBP13 were indeed closely related to the

Pbprp2/Pbprp5 of *Drosophila*. CquiOBP55 and AegOBP83 identified in this analysis are orthologues of AgamOBP29 and homologous to OBP58 of *Drosophila*. Interestingly three orthologous sequences between the three species were found closely related to *MinusC* members of the *Drosophila*. The other clusters identified are named as *mclassic1-9* (Figure 5a and supplementary Figure 2a). In addition one cluster with 16 members lacking C2 and C5 cysteines has been named as *Bombyx mori MinusC* (Zhou et al. 2009) due to their homology with the *B. mori* sequences respectively. This homology was determined using blast analysis and confirmed with the *inParanoid* eukaryotic ortholog database (O'Brien et al. 2008). Other subtype classifications of the *Classic* members were also similar to the clustering seen in the *inParanoid* database. The *Classic* OBPs from *Aedes aegypti* and *Culex quinquefasciatus* shared orthologues with *Anopheles gambiae* that were overwhelmingly mapped to the species chromosome 2R. The *Bombyx mori Minus C* proteins shared preferred orthology with OBPs that were mapped to chromosome 3R in *Anopheles gambiae*. The sequences homologous to the *Drosophila Minus C* shared orthology with the corresponding AgamOBP that is mapped to the X chromosome. It is interesting to note that two OBP genes from *Anopheles gambiae* homologous to the *Drosophila LUSH* were located on two different chromosomes 3R and 3L (supplementary Figure 1d and 1e). Similarly, AgamOBP20 from the Obp19 subtype is located on chromosome 2L while the other members of this subtype locate themselves on chromosome 2R. These and their corresponding orthologues in *Aedes* and *Culex* are shown in supplementary Figure 1b and 1c. It was also observed that most of the OBP genes homologous to the pheromone binding protein types of the *Drosophila* genome shared orthology/paralogy with the corresponding genes located on the chromosome 2R in *Anopheles*.

The *PlusC* OBPs clustered as 9 major clusters forming 9 subtypes (*mplus1-mplus9*). However it was difficult to interpret the molecular background behind this clustering. Interestingly, except for *mplus9* members which localized on chromosome 3L, all the *PlusC* OBPs from *Anopheles gambiae* were distributed on chromosome 2L which harbors, in addition, one *Atypical* and one *Classic* member.

2.3.7. Sequence specific clustering of *Atypical* odorant binding proteins

The *Atypical* OBPs unlike the *Classic* members formed just four major clusters which are named in this study *matype1-matype4* and the clustering is characteristic of their sequence features. The *matype1* forms the smallest cluster among the 4 subtypes with two members from each genome and this cluster is separated from the other three subtypes with high bootstrap values. Interestingly

these members were closely related to the *Classic* members when observed on the phylogeny of individual genomes. The *matype2* forms a distinctive type of *Atypical* members holding only a total of 6 cysteines (C1, C3, C4, C5, C1', C6') out of the 12 conserved cysteines characteristic of the other subtypes of this subfamily (Figure 1). This interesting conservation of 12 cysteines throws lights on the evolution of *Atypical* members and opening a new dimension towards its unique features. The extended C terminal of these members with the increase in the number of members identified in this chapter described above highlighted the presence of 6 cysteines (C1'-C6') and they hold a similar cysteine spacing as observed in the N-terminal end. This striking feature validates the hypothesis that members in this class of proteins are indeed two domains proteins but yet distantly related to the *Classic* OBPs. When the two domains were analysed separately for their homologous sequences, the search always identified *Classic* OBP members as closest members (Table 2). This confirmed the fact that these proteins are indeed two domain proteins whose features were not recognised prior to this work and the work of Vieira and Rozas (2011), mainly due to limited members identified in this subfamily. In this study, this subfamily has been extended with 57 new members. The *matype2* still features to stand as a distinctive type with the presence of cysteines in the N terminal domain lacking C2 – C5 encouraging them to be called *Minus -C* like *Atypical* proteins. The *matype4* members unanimously hold a deletion of about 15 residues between the C1 and C2 which stands as the distinguishing feature of this subtype. The *matype1* members are orthologous to AgamOBP39 that is located on chromosome 2R which is otherwise populated with *Classic* members supporting their close relation to the *Classic* members in the phylogeny of the individual genomes. The *matype2* members intriguingly share orthology with corresponding OBPs from *Anopheles gambiae* that were mapped to chromosome X whereas *matype3* and *matype4* members were sharing orthology with AgamOBPs distributed over chromosomes 3R and 3L.

2.4. Discussion

2.4.1. Rapid evolutionary based duplication in the *Culicinae* family of mosquitoes

The *Culex quinquefasciatus* genome (Arensburger et al. 2010) and the *Aedes aegypti* genome (Nene et al. 2007) code for 109 and 111 OBPs respectively with the *Anopheles* genome coding for only 67 OBPs. This is evident with the increase in the genome size that has been observed in the two species with respect to the gene duplication events of important genes involved in the adaptation to the environment. These putative OBPs identified in the three species fall into three

major subfamilies the *Classic*, *PlusC* and *Atypical* described based on their sequence features in comparison to the OBP subfamilies in *Drosophila melanogaster* (Xu et al. 2003; Zhou et al. 2008; Pelletier and Leal 2009). The identification of the 26 *Atypical* and 12 *PlusC* OBP members in the *Culex quinquefasciatus* genome and a remarkable expansion of *Atypical* OBPs in the *Aedes aegypti* genome with 31 additional members is reported which has opened insights for a revised classification of OBPs in the mosquito genomes discussed further in this chapter. In general, 61 new OBPs in *Culex*, 47 new OBPs in *Aedes* and 3 OBPs in *Anopheles* have been identified. The increase in the number of members identified is a reflection of the careful examination of sequences for cysteine conservation patterns. Based on their orthology with the genes located on the *Anopheles gambiae* chromosomes, putative chromosomal mapping of *Aedes* and *Culex* OBP genes could provide a picture of their distribution in these genomes, though the exact chromosomal location of the supercontigs from these two species has yet to be established by physical mapping. It is however clear from the current data that many of the duplicated genes in the *Culex* and *Aedes* species appeared as gene clusters. Many of the *Classic* AgamOBP members are primarily located on chromosome 2R while the *PlusC* AgamOBP members are found on the chromosome 2L. The *Atypical* members are distributed evenly in all the chromosomes with *matype2* which forms a distinctive cluster in the *Atypical* subfamily housed in the X chromosome.

2.4.2. Functional sub clustering of the odorant binding proteins

The comparative analysis of the three main subfamilies *Classic*, *Plus C* and *Atypical* among the three mosquito species and the *Drosophila* OBPs in the case of *Classic* OBPs indicated the extensive diversity among each subfamily. The use of structure alignment for the construction of alignments helped in retaining important sequence features in turn improving the resolution of the phylogenetic trees. The use of the distantly related D7 family members as an outgroup increased the fidelity of the branching patterns leading to more reliable clustering of the diverse sequences. The *Classic* OBP subfamily holds 19 subtypes few of which were named previously (Xu et al. 2003; Zhou et al. 2008; Pelletier and Leal 2009) as Pbprp1, Pbprp4, Lush, OSE/OSF and OBP19a based on their homology with *Drosophila* OBPs. Similarly members closely related to Pbprp2/Pbprp5, OBP58 and OBP99c of *Drosophila* were identified and named them accordingly. In addition members closely related to the *Minus C* proteins of *Bombyx mori* have been identified and named as *Bombyx mori Minus C* proteins. 9 new clusters have been identified and named as *mclassic1* –

mclassic9 (mosquito *Classic*) based of their clustering patterns. These members do not share considerable homology with the *Drosophila* OBP members.

The clustering of the *Atypical* odorant binding proteins revealed the presence of four major subtypes *matype1-matype4* (Mosquito *Atypical*) with observable common and distinct sequence features between the different subtypes. All the members of *matype1*, *matype3*, *matype4*, with a few exceptions, carried 12 conserved cysteines named C1-C6 and C1'-C6'. This is the first detailed report of such an observation while the previous analysis of this subfamily just indicated the presence of an extended C-terminal end with unknown features or its homology to *Classic* OBP domains (Vieira and Rozas 2011). The *matype2* carries 6 cysteines aligned to C1, C3, C4, C6, C4' and C6'. The *matype4* was found to have a deletion of about 15 residues which are retained in the other three subtypes. The *matype1* was more closely related to the AgamOBP39 which is evident in the phylogeny of the individual genomes and they were indeed found in close proximity to the *Classic* OBPs, at least in the case of the *Anopheles gambiae* genome. The *PlusC* OBPs form 9 novel major clusters and are named as *mplus1-mplus9* subtypes. They similarly to the *Atypical subfamily*, do hold recognizable sequence features (supplementary Figure 2c).

2.4.3. *Atypical* OBPs are indeed Two-Domain OBPs

The increase in the number of *Atypical* OBPs in the three mosquito genome revealed important facets in this subfamily of proteins. A total of 57 new members were added to this subfamily which represents more than a two fold increase than the previously identified proteins 29 members (Xu et al. 2003; Zhou et al. 2008). The *Atypical* members were first identified in the *Anopheles gambiae* genome and were described as larger proteins holding same conserved cysteines as the *Classic* OBPs in their N-term region and that, in addition, have a characteristic extended C-terminal end. The C-terminal however, in the current analysis, was found to hold a repeat in the conservation of the C-terminal cysteines which was not previously reported mainly due to the smaller number of members identified in this subfamily. The cysteines in the C-terminal extension have been named C1'-C6' accordingly. This remarkable conservation of cysteines is believed to hold important evolutionary information. Further analysis of the N-terminal and C-terminal domains of these protein separately using blast analysis revealed the identification of *Classic* OBP members by each of these domains with often significant E-values raising the curiosity that the members of these family are indeed two-domain OBPs thus confirming the preliminary observation of Vieira and Rozas (2011). Interestingly the *Classic* OBP members obtained as hits by

each of these domain were found to be closely related to the *Minus C* family of proteins in the *Drosophila* genome as observed in their known “Dimer OBPs” 83cd and 83ef (Zhou et al. 2004) which are proteins that hold two OBP domains. It is also noticed in this study that the two domains often picked up two different *Classic* members suggestive of heterologous combination events. It could be speculated that these proteins evolved in significance to the reduction of cell cost which could otherwise be used in the formation of functional dimers. The recent publication of a functional dimer in the *Culex quinquefasciatus* genome (Mao et al. 2010) supports the current important speculations on *Atypical* members indicating the importance of the presence of two domain proteins in the binding of relatively large ligands. Thus it is confirmed that the *Atypical* OBP members are indeed two-domain OBPs which are previously observed in *Drosophila* as “Dimer OBPs” and that they do not stand specific to the mosquito genomes as reported earlier (Xu et al. 2003). Furthermore the *matype2* members which carry a presence of only 6 cysteines in the place of 12 cysteines in the other *Atypical* types is suggestive of a possible adaptation in the fold with three disulphide bonds in place of 6 disulphide bonds in the other types. The astound putative distribution of these genes in the X-chromosome further increases the speculative importance of these proteins in the blood feeding mechanism by female mosquitoes and stand as a very important finding in the current analysis. Overall the structural determination and ligand binding studies of the members of the *Atypical* members which is proposed here to be called *Two-domain* OBP proteins would be of significant importance in deciphering the olfactory mechanism in the mosquito species.

2.4.4. *Minus C* proteins in the mosquito genomes.

The *Minus C* subfamily of OBPs was first identified in the *Drosophila* genome with 7 members with some of its members lacking the second and fifth cysteine residues which encouraged the naming of this subfamily as *MinusC*. However some proteins retained six cysteines as it is the case for members from *Anopheles gambiae*. These members were retained as a part of this cluster based on the alignment data but was also because they appeared as gene clusters. The *MinusC* proteins in the mosquito species have not been described previously but the current analysis reveals the clustering of three orthologous OBP sequences AgamOBP9, AegOBP22 and CquiOBP43 with the *Drosophila Minus C* members OBP99a, OBP44a and OBP99b (Figure 5a) with a considerable bootstrap support among which OBP99a alone retains all the six cysteines. The mosquito sequences however retain all the six cysteines. This cluster has been named OBP99a (*Minus C*). In addition to this cluster the *Culex quinquefasciatus* genome was found to hold a cluster

of 15 members (Figure 4c) all of which lacked the C2 and C5 cysteines, but intriguingly, they were not closely related to the *Drosophila Minus C* subfamily members. Further analysis of this cluster using BLAST and with reference to the clustering available in the inParanoid database for eukaryotic genomes (O'Brien et al. 2005), it was found that these proteins were closely related to the *Bombyx mori Minus C* proteins. This cluster has been named *Bombyx mori Minus C* in relation to their homology with the *Bombyx mori Minus C* proteins. However the other two genomes lack this cluster with *Aedes aegypti* carrying just one member closely related to this class of proteins. It is interesting to note that 10 of these members (supplementary Figure 1e) appear as a single gene cluster on supercontig3.26 in the *Culex* which is mapped orthologues on chromosome 3R in *Anopheles gambiae* suggestive of gene duplication events required for the adaptation to environmental cues. The members of the *Minus C* subfamily of proteins thus stand important candidates for further analysis both on structural and functional aspects.

2.5. Conclusion

The current analysis provides a massive expansion of odorant binding proteins with a total of 113 members in the three mosquito genomes *Anopheles gambiae*(4), *Aedes aegypti*(47), *Culex quinquefasciatus*(61). The current expansion has helped in the in-depth characterization of the various subtypes within the *Classic*, *Plus C* and *Atypical* subfamilies of OBPs. It stands as the first detailed analysis reporting the existence of “Dimer” /two-domain OBPs and Minus C OBPs in the mosquito genomes. It also reports the identification of a unique subtype among the “Dimer/two-domain OBPs which in the current analysis is called the ‘*Minus-C like Atypical OBPs*’ .

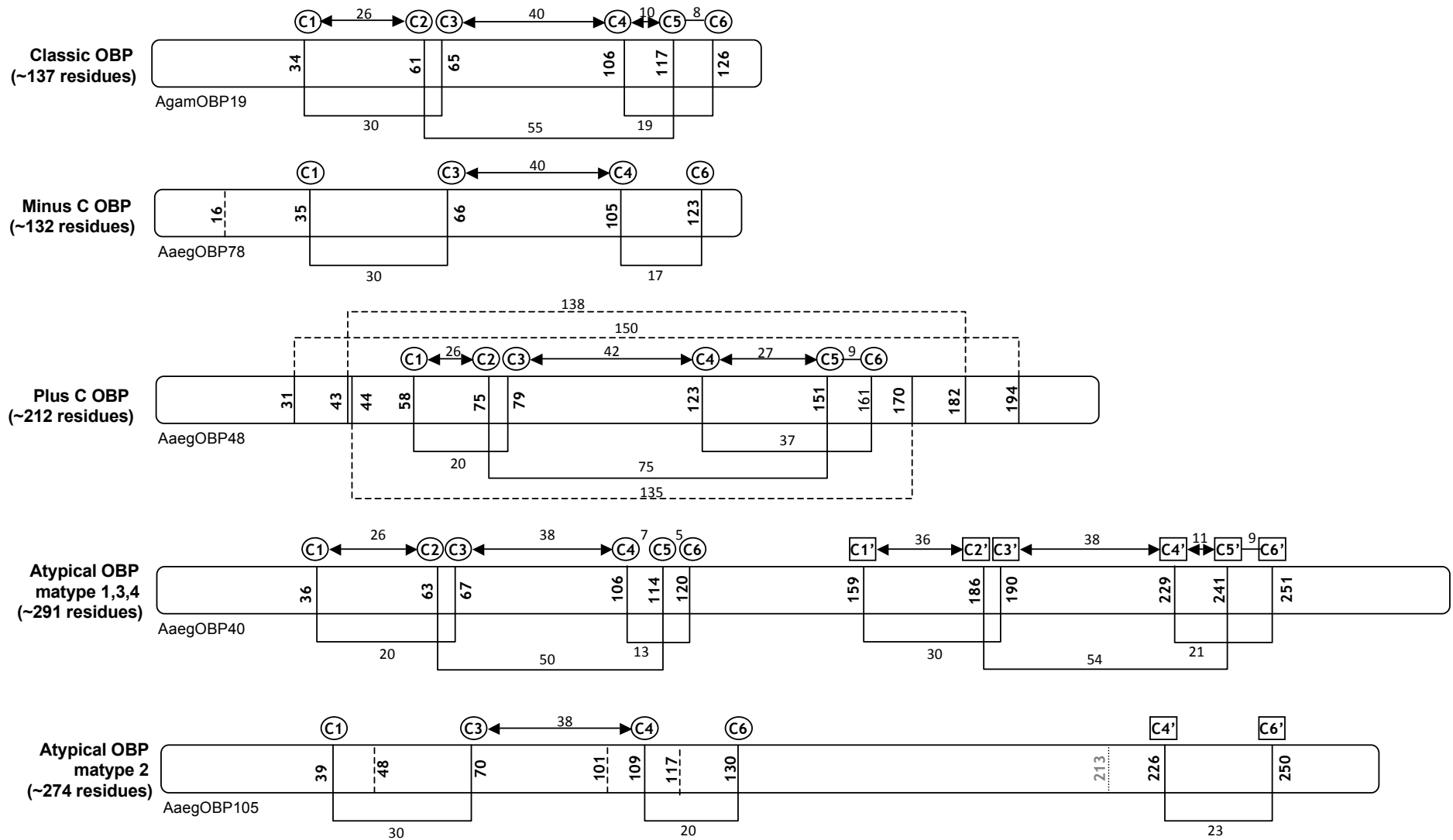
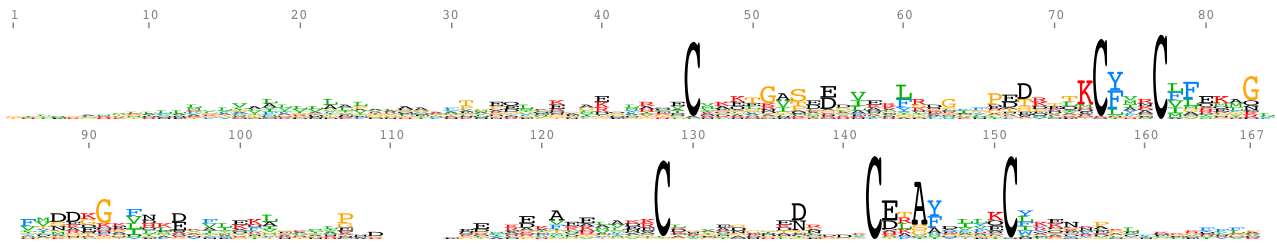
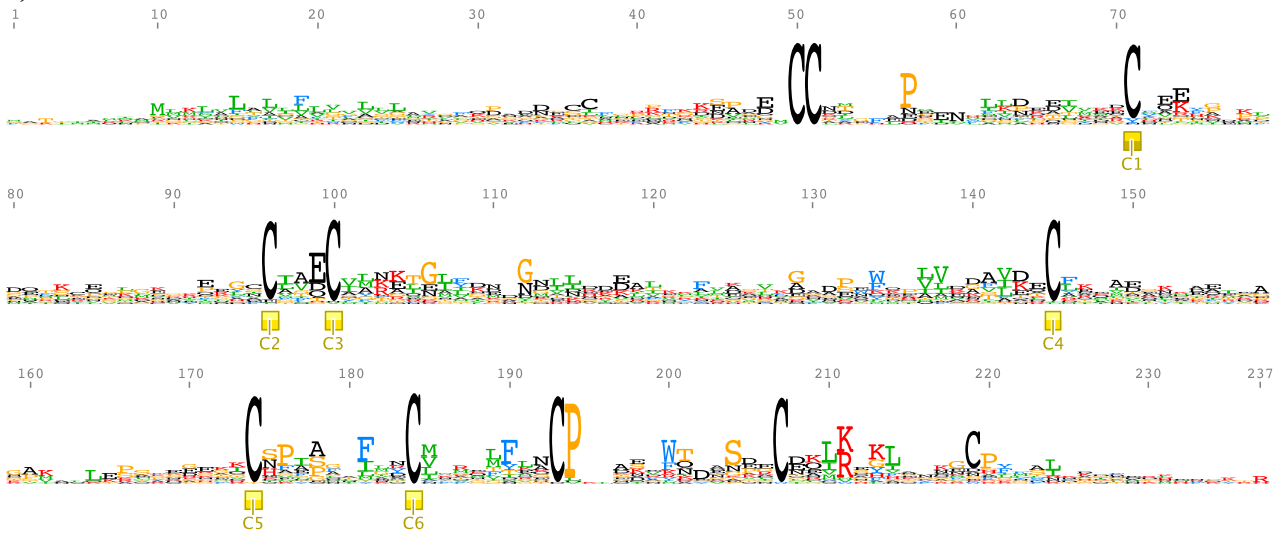


Figure 2.1. Cysteine conservation patterns across the different subfamilies and subgroups of odorant binding proteins from *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* genomes. The 6 conserved cysteines in GOBP domain are denoted C1-C6. The 6 additional cysteines in the C-term of the *Atypical* OBPs are denoted C1'-C6'.

a) Classic OBPs



b) PlusC OBPs



c) Atypical OBPs

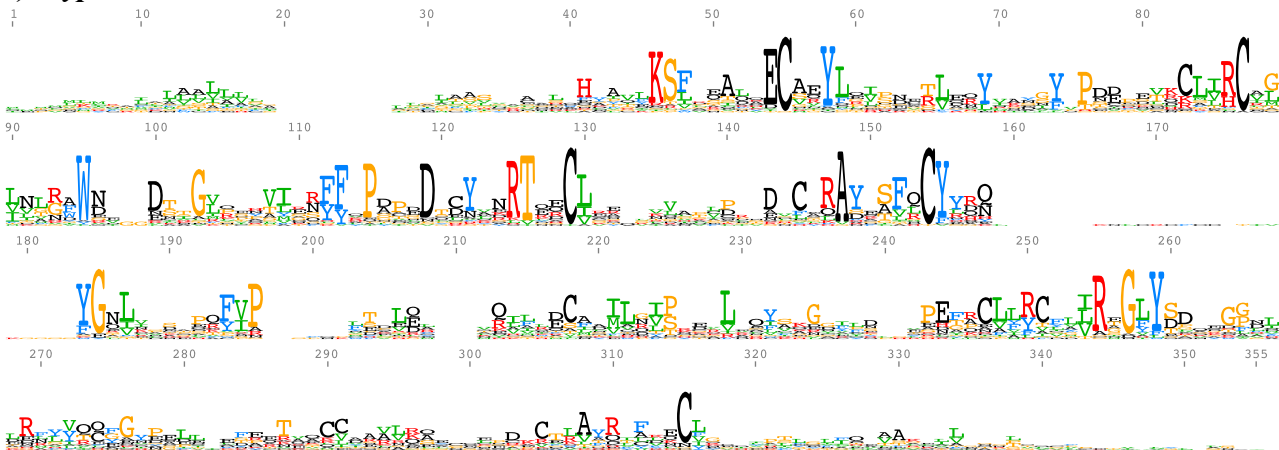


Figure 2.2. Residue conservation patterns within each OBP subfamily from *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* in the form of sequence logos. Detailed sequence alignments for each cluster from these subfamilies are provided in supplementary Figures 2a-c.

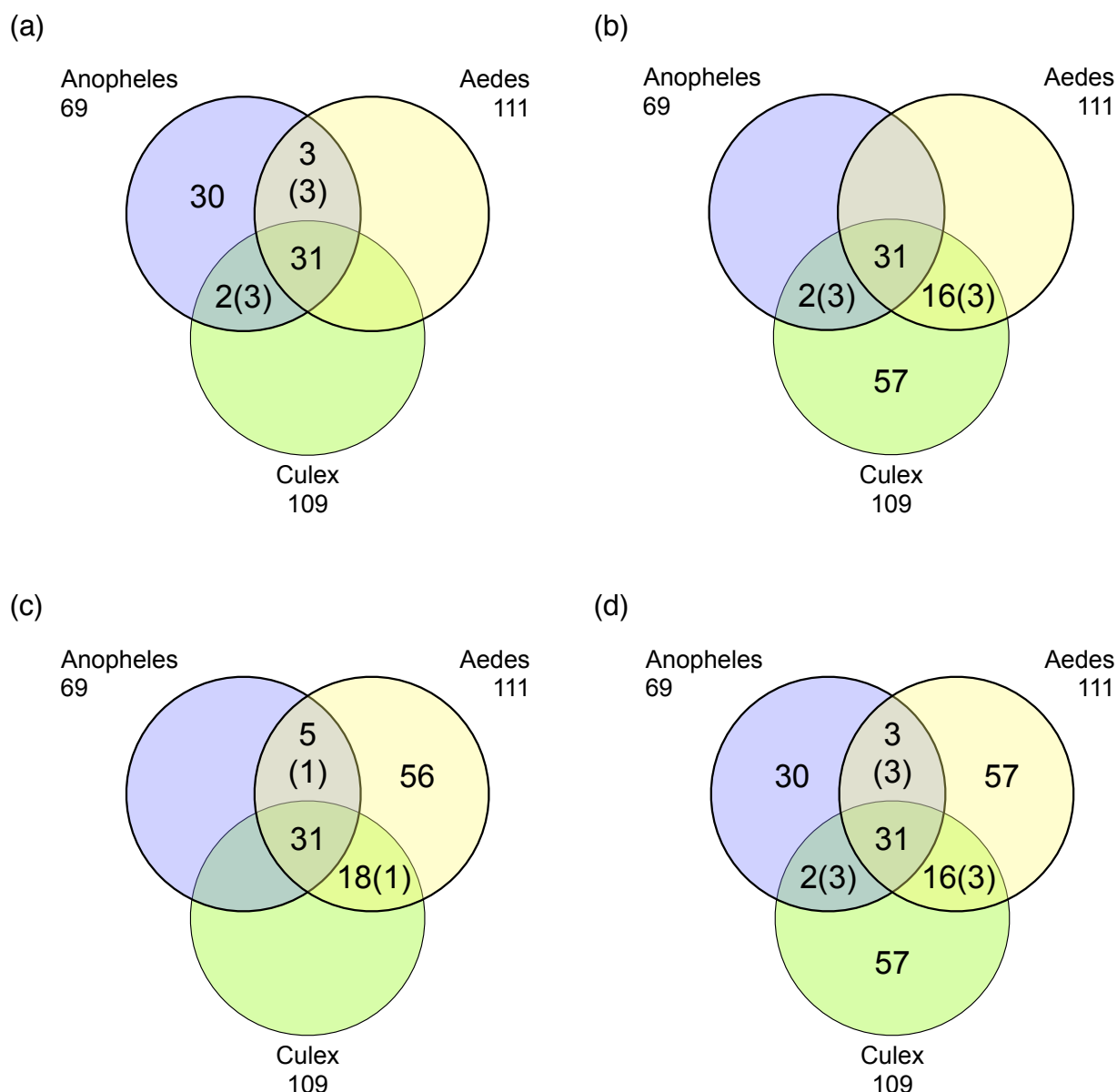


Figure 2.3. Analysis of orthologous OBP genes shared across three mosquito species, *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*. The Venn diagrams indicate the number of inferred orthologous genes shared among the mosquitoes species : (a) number of *A. gambiae* OBP genes orthologous to *A. aegypti* and *C. quinquefasciatus*; (b) number of *A. aegypti* OBP genes orthologous to *A. gambiae* and *C. quinquefasciatus*; (c) number of *Culex* OBP genes orthologous to *A. gambiae* and *A. aegypti* ; (d) overall number of orthologous groups across the three mosquito species. The number of genes that share a three way (1:1:1) orthology between the 3 species is 31. The number of genes in a species that have two way orthology (1:1) with the two other species but not a three way orthology is indicated between parenthesis and for a given species, should be counted only once. For example, in 3(a), the total number of OBP genes in *Anopheles gambiae* is $30 + 3 + 2 + 31 + (3) = 69$ since 3 genes in *A. gambiae* have two way orthology (1:1) with genes in both *C. quinquefasciatus* and *A. aegypti* but not a three way orthology. Detailed listings of the orthology analysis are provided in supplementary Tables 1a, 1c and 1d.

Anopheles gambiae

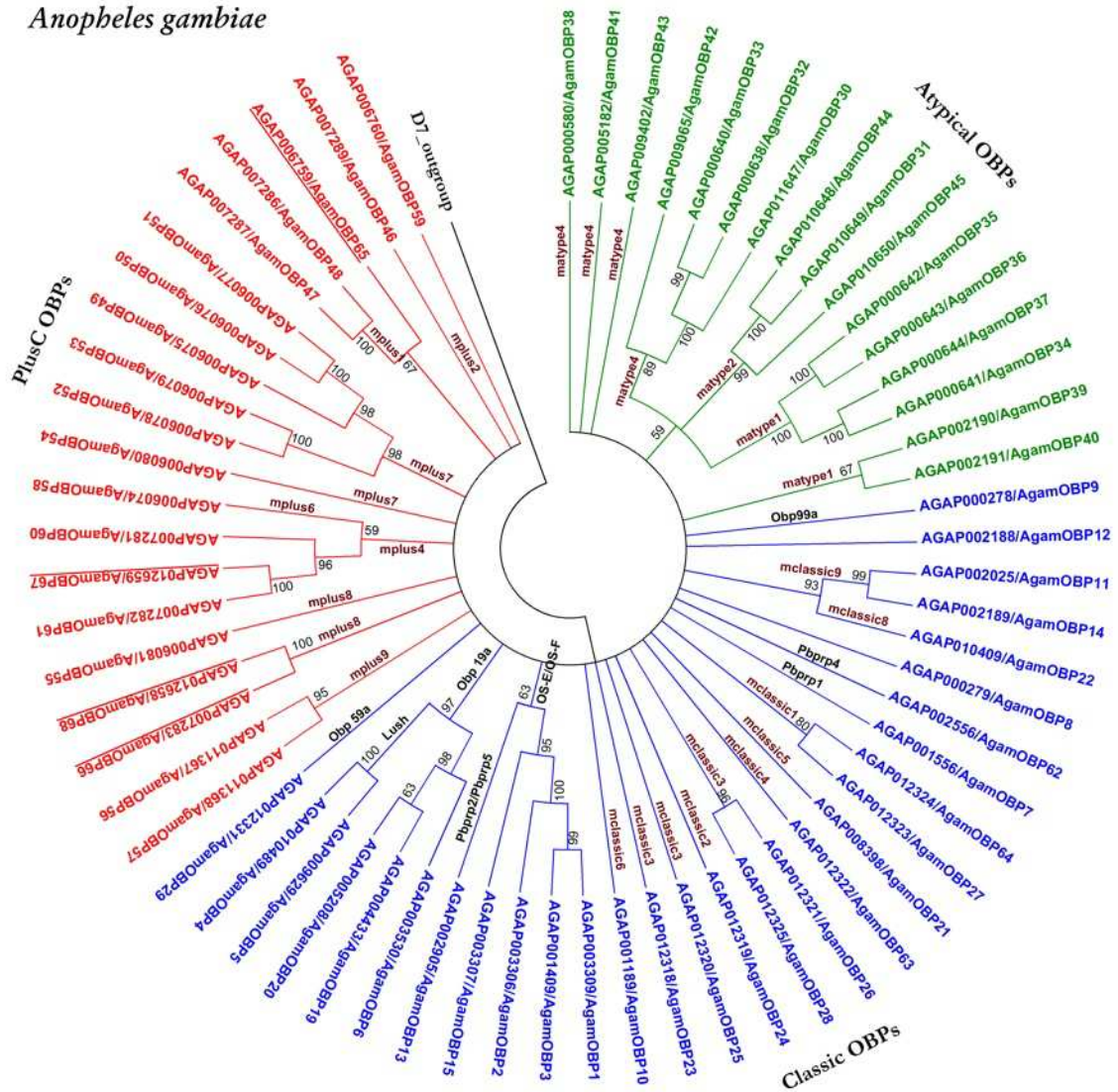


Figure 2.4a. Rooted phylogenetic tree of the odorant binding proteins in the *Anopheles gambiae* genome. The *Classic* OBPs subfamily are colored blue, *Atypical* OBPs are colored green and *PlusC* OBPs are colored red. The bootstrap values are indicated on the nodes in percentage values. The names of identified clusters are indicated on the branches (see Figure 5).

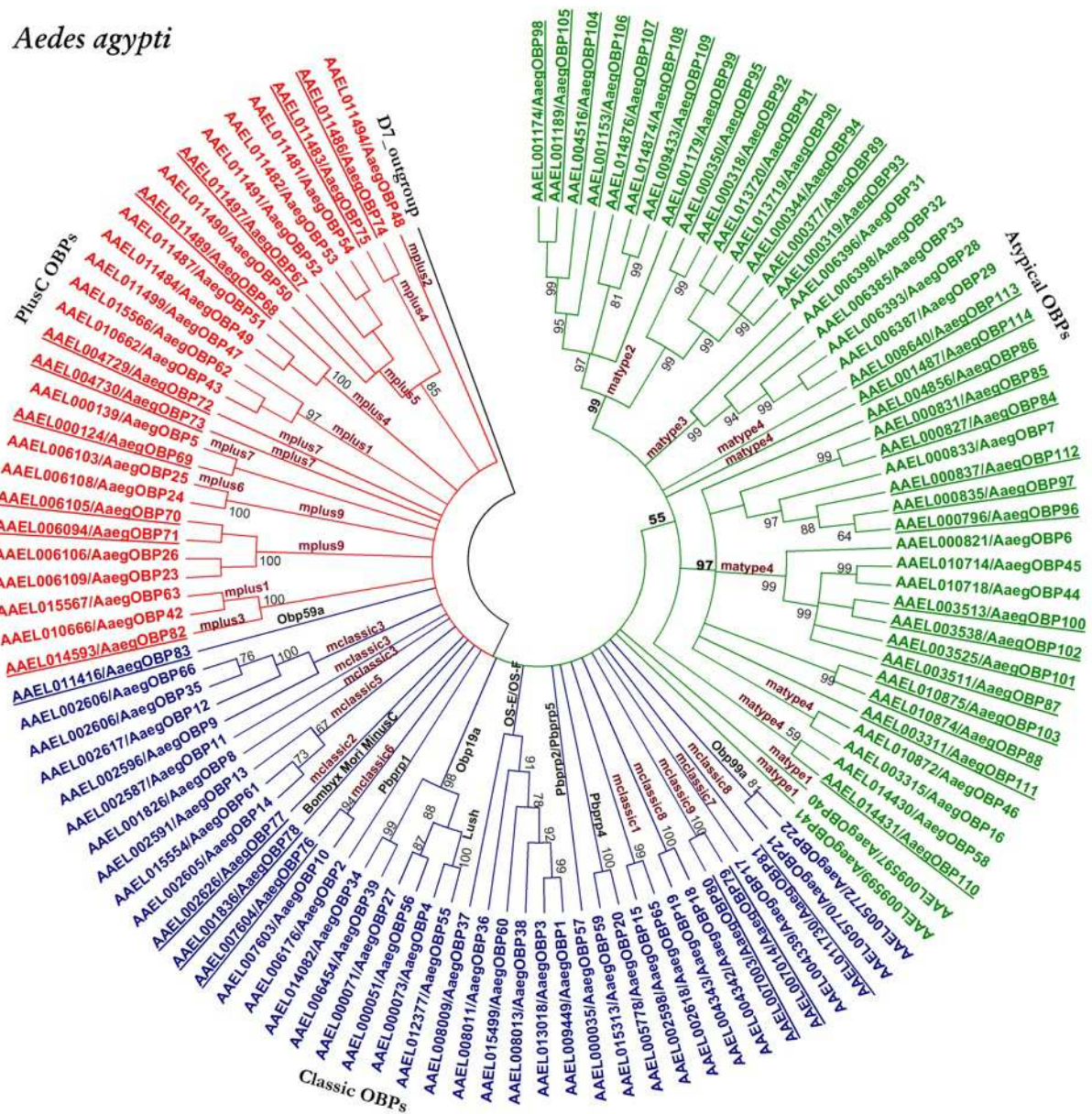
Aedes aegypti

Figure 2.4b. Rooted phylogenetic tree of the odorant binding proteins in the *Aedes aegypti* genome. The *Classic* OBPs subfamily are colored blue, *Atypical* OBPs are colored green and *PlusC* OBPs are colored red. The bootstrap values are indicated on the nodes in percentage values. The names of identified clusters are indicated on the branches (see Figure 5).

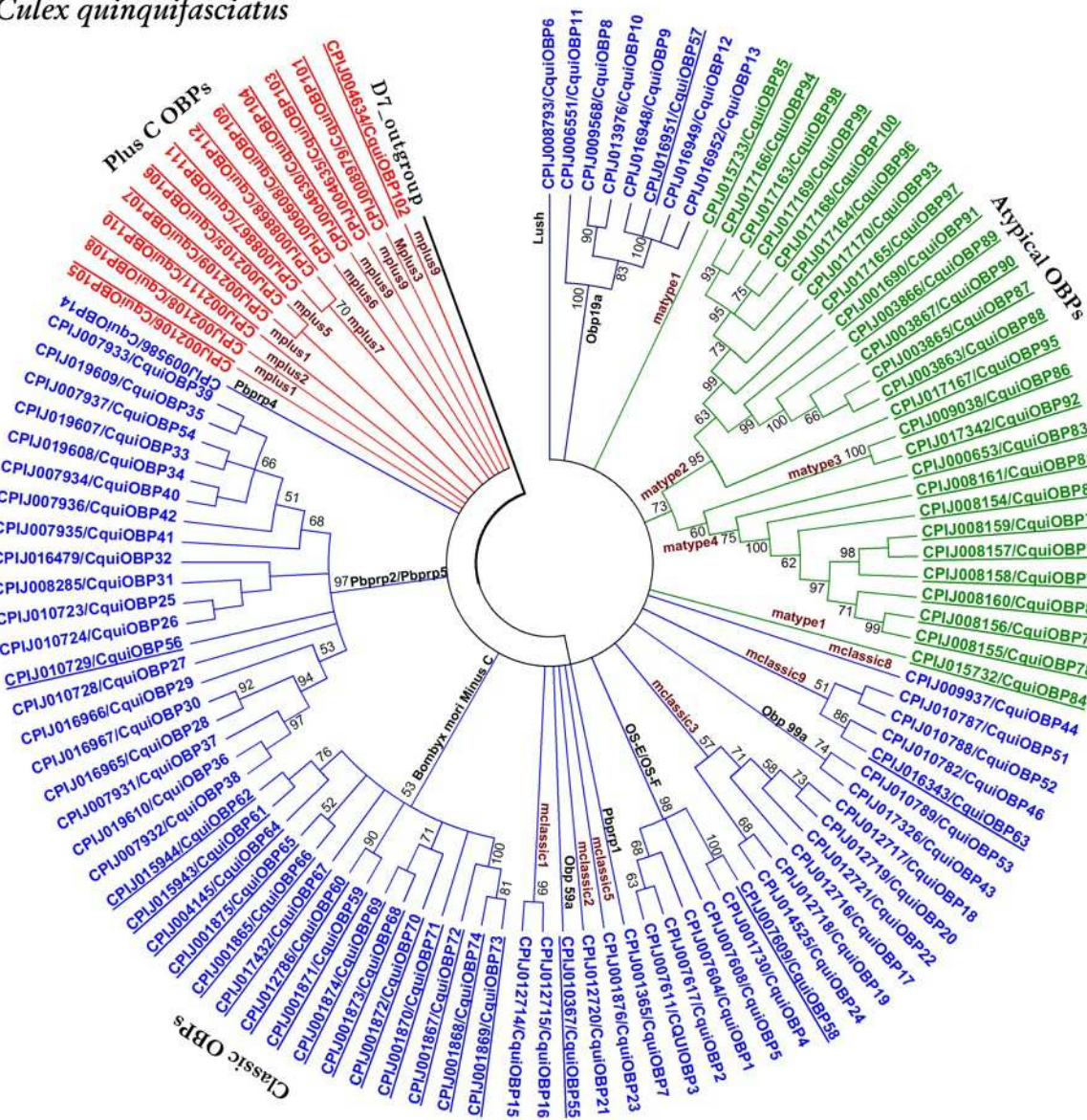


Figure 2.4c. Rooted phylogenetic tree of the odorant binding proteins in the *Culex quinquefasciatus* genome. The *Classic* OBPs subfamily are colored blue, *Atypical* OBPs are colored green and *PlusC* OBPs are colored red. The bootstrap values are indicated on the nodes in percentage values. The names of identified clusters are indicated on the branches (see Figure 5).

52

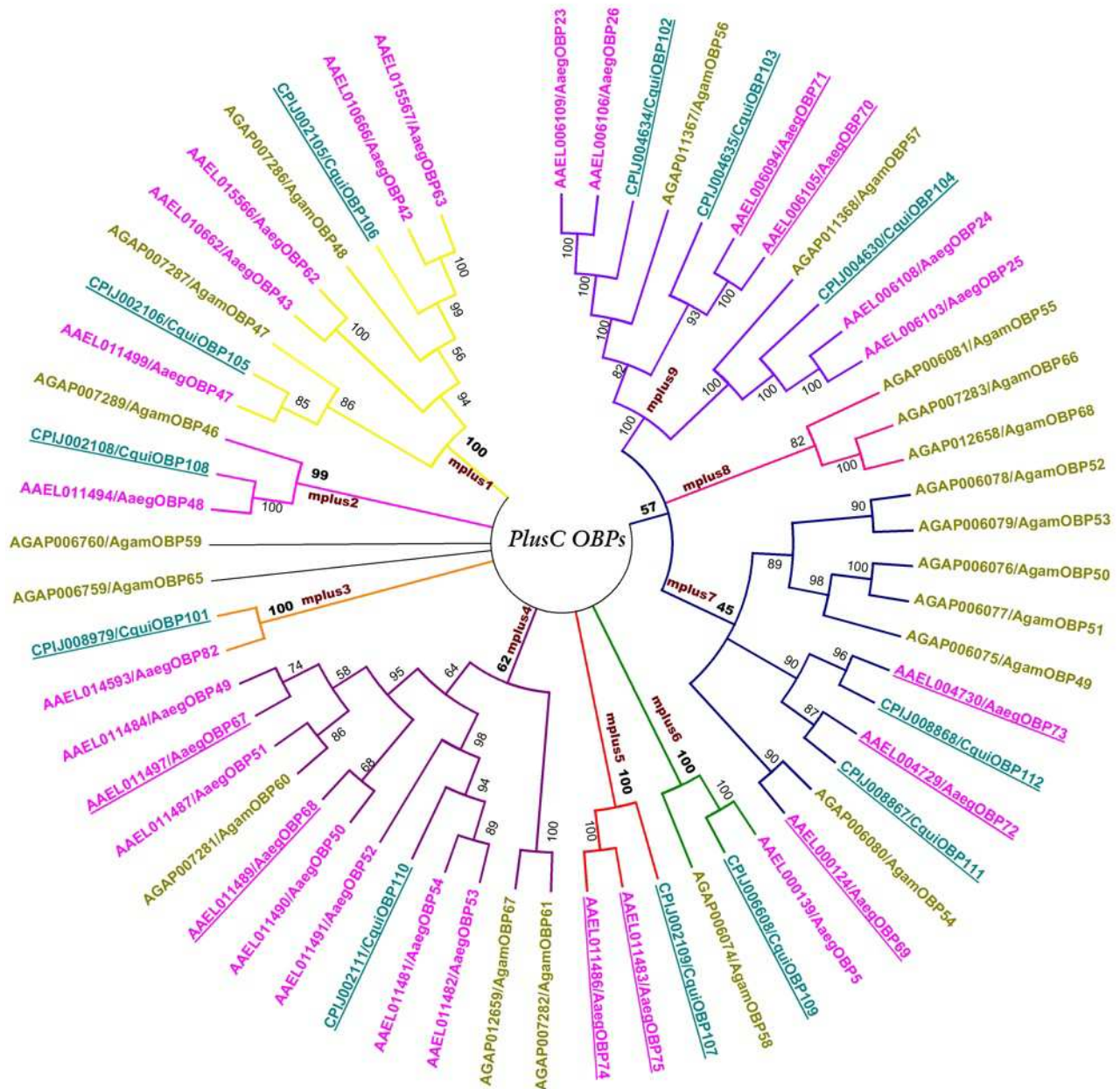


Figure 2.5b. Unrooted phylogenetic tree of *PlusC* odorant binding proteins in the three mosquito genomes. The *An. gambiae*, *Ae. aegypti* and *C. quinquefasciatus* members are colored in mustard, pink and turquoise respectively. The bootstrap values are indicated on the nodes in percentage values. The names of identified clusters inside the *PlusC* OBPs subfamily are indicated on the branches. Detailed alignments of the members inside each cluster are provided in supplementary Figures 2b.

Figure 2.5c. Unrooted phylogenetic tree of *Atypical* odorant binding proteins in the three mosquito genomes. The *An. gambiae*, *Ae. aegypti* and *C. quinquefasciatus* members are colored in mustard, pink and turquoise respectively. The bootstrap values are indicated on the nodes in percentage values. The names of identified clusters inside the *Atypical* OBPs subfamily are indicated on the branches. Detailed alignments of the members inside each cluster are provided in supplementary Figure 2c.

Table 2.1. Identification of OBPs in *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* genomes. Are shown the statistics of previously and newly identified OBP members (AgamOBP65 to AgamOBP68, AaegOBP67 to AaegOBP114, CquiOBP54 to CquiOBP112) in all three mosquito genomes. Detailed results are provided in accompanying supplementary Tables 1a-d.

		Sub family				New total
		Classic	Plus C	Atypical	not determ.	
<i>Anopheles gambiae</i>	Previously reported ¹	29	16	16		69
	newly identified		4		4	
<i>Aedes aegypti</i>	Previously reported ²	33	17	14		111
	newly identified	6	10	31		
<i>Culex quinquefasciatus</i>	Previously reported ³	48				109
	newly identified	21	12	26	2	

¹ - Vogt et al. 2002; Xu et al. 2003; Zhou et al. 2004; Vieira and Rozas 2011.

² - Zhou et al. 2008 Pelletier et al. 2010.

³ - Pelletier et al. 2010.

Table 2.2. Analysis of the two putative OBP domains (N-term and C-term) of *Atypical* OBPs from *An. gambiae*, *Ae. aegypti* and *C. quinquefasciatus*. Shown are the results of the Psi-Blast search among all mosquito *Classic* OBPs and *Drosophila* OBPs after splitting the *Atypical* proteins into their two respective putative domains.

Mosquito Atypical OBP		Mosquito classic OBP closest homologues						Drosophila OBP closest homologues			
ID	Phylogenetic subgroup	N-term ¹	Phylogenetic subgroup	E-value	C-term ²	Phylogenetic subgroup	E-value	N-term ¹	E-value	C-term ²	E-value
AGAP000638* AgamOBP32	-	AGAP010409 AgamOBP22	mclassic8	1.84E-10	AGAP002025 AgamOBP11	mclassic9b	2.38E-06	OBP99a	3.15E-06	OBP99c	3.90E-03
AGAP009402* AgamOBP43	-	AGAP010409 AgamOBP22	mclassic8	3.93E-10	AGAP002189 AgamOBP13	mclassic9b	1.24E-05	OBP99a	1.17E-05	OBP99a	6.33E-09
AAEL003538* AaegOBP102	matype4	AAEL011730 AaegOBP81	mclassic8	6.30E-06	AAEL005770 AaegOBP21	-	2.08E-07	OBP99d	1.44E-01	OBP99a	7.26E-05
AAEL003511* AaegOBP87	matype4	AAEL011730 AaegOBP81	mclassic8	3.16E-03	AAEL005770 AaegOBP21	-	5.83E-10			OBP99a	4.47E-05
AAEL003513* AaegOBP100	matype4	AAEL011730 AaegOBP81	mclassic8	5.50E-06	AAEL005770 AaegOBP21	-	2.89E-07	OBP99d	1.40E-01	OBP99a	7.18E-05
AAEL010718* AaegOBP44	matype4	AAEL011730 AaegOBP81	mclassic8	5.60E-06	AAEL005770 AaegOBP21	-	3.97E-06	OBP56g	1.18E-02	OBP99a	3.17E-04
AAEL003311 AaegOBP111	matype4	AAEL011730 AaegOBP81	mclassic8	2.11E-04	AAEL002596 AaegOBP9	mclassic3b	6.20E-04			OBP99d	8.97E-03
AAEL004856 AaegOBP86	matype4	AAEL011730 AaegOBP81	mclassic8	1.52E-04	AAEL007014 AaegOBP79	-	1.15E-05	OBP56i	3.02E-03	OBP99a	1.63E-04
AAEL000831 AaegOBP85	matype4	AAEL011730 AaegOBP81	mclassic8	2.04E-02	AAEL002596 AaegOBP9	mclassic3b	6.14E-05			OBP56g	3.21E-04
AAEL000821 AaegOBP6	matype4	AAEL011730 AaegOBP81	mclassic8	4.19E-05	AAEL005770 AaegOBP21	-	2.90E-06	OBP99d	1.40E-02	OBP99a	9.17E-02
AAEL010872 AaegOBP46	matype4	AAEL011730 AaegOBP81	mclassic8	1.05E-03	AAEL004342 AaegOBP18	mclassic9a	1.62E-04			OBP99b	6.20E-03
AAEL000827 AaegOBP84	matype4	AAEL007014 AaegOBP79	-	6.49E-03	AAEL005770 AaegOBP21	-	1.98E-02	OBP99b	8.14E-01	OBP99a	1.90E-04
AAEL000835 AaegOBP97	matype4	AAEL011730 AaegOBP81	mclassic8	3.57E-02	AAEL004343 AaegOBP19	mclassic9a	1.89E-06	OBP57b	2.73E+00	OBP56i	5.28E-05
AAEL000796 AaegOBP96	matype4	AAEL011730 AaegOBP81	mclassic8	3.10E-03	AAEL004343 AaegOBP19	mclassic9a	7.52E-06	OBP57b	2.62E+00	OBP56i	2.50E-04
AAEL010874 AaegOBP88	matype4	AAEL011730 AaegOBP81	mclassic8	2.41E-05	AAEL005770 AaegOBP21	-	3.75E-06	OBP56g	4.56E-01	OBP99d	2.35E-04
AAEL010875 AaegOBP103	matype4	AAEL011730 AaegOBP81	mclassic8	2.18E-05	AAEL005770 AaegOBP21	-	4.40E-05	OBP56g	3.60E-01	OBP99a	2.70E-04
AAEL010714 AaegOBP45	matype4	AAEL011730 AaegOBP81	mclassic8	3.17E-04	AAEL005770 AaegOBP21	-	3.03E-07	OBP56g	3.14E-02	OBP99a	1.95E-05
AAEL003525 AaegOBP101	matype4	AAEL011730 AaegOBP81	mclassic8	4.07E-03	AAEL005770 AaegOBP21	-	6.78E-07			OBP99a	4.20E-04
AAEL000833 AaegOBP7	matype4	AAEL007003 AaegOBP80	-	9.85E-03	AAEL004343 AaegOBP19	mclassic9a	2.51E-04	OBP19b	1.72E+00	OBP99d	2.13E-04
AAEL003315 AaegOBP16	matype4	AAEL011730 AaegOBP81	mclassic8	2.28E-08	AAEL005770 AaegOBP21	-	6.50E-04	OBP56d	1.23E+00	OBP99c	1.03E-02
CPIJ008158* CquiOBP76	matype4	CPIJ010789 CquiOBP53	mclassic7	7.20E-02	CPIJ016343 CquiOBP63	mclassic9b	4.60E-10	OBP56e	2.82E+00	OBP99a	6.38E-07
CPIJ008159* CquiOBP77	matype4	CPIJ009937 CquiOBP44	mclassic8	2.58E-02	CPIJ016343 CquiOBP63	mclassic9b	5.93E-10	OBP99d	7.90E+00	OBP99a	2.43E-08
CPIJ008155* CquiOBP78	matype4	CPIJ009937 CquiOBP44	mclassic8	1.97E-02	CPIJ016343 CquiOBP63	mclassic9b	2.58E-06	OBP99d	6.77E-01	OBP99a	4.05E-04

Mosquito Atypical OBP		Mosquito classic OBP closest homologues						Drosophila OBP closest homologues			
ID	Phylogenetic subgroup	N-term ¹	Phylogenetic subgroup	E-value	C-term ²	Phylogenetic subgroup	E-value	N-term ¹	E-value	C-term ²	E-value
CPIJ008154 CquiOBP82	matype4	CPIJ014525 CquiOBP24	mclassic6	2.14E-02	CPIJ016343 CquiOBP63	mclassic9b	1.37E-04	OBP56g	4.21E-02	OBP99a	2.49E-02
CPIJ008161 CquiOBP81	matype4	CPIJ009937 CquiOBP44	mclassic8	4.16E+00	CPIJ016343 CquiOBP63	mclassic9b	2.53E-09			OBP99b	6.95E-06
CPIJ008157 CquiOBP75	matype4	CPIJ0010787 CquiOBP51	mclassic9a	7.87E-02	CPIJ010782 CquiOBP46	mclassic9b	2.80E-09			OBP99a	2.08E-07
CPIJ000653 CquiOBP83	matype4	CPIJ009937 CquiOBP44	mclassic8	1.70E-01	CPIJ016343 CquiOBP63	mclassic9b	4.70E-06	OBP57c	2.57E+00	OBP99b	1.23E-03
CPIJ008156 CquiOBP79	matype4	CPIJ010782 QuiOBP46	mclassic9b	2.06E-01	CPIJ010782 CquiOBP46	mclassic9b	5.29E-11	OBP99d	1.52E+00	OBP99a	6.64E-08
CPIJ008160 CquiOBP80	matype4	CPIJ010789 CquiOBP53	mclassic7	8.70E-02	CPIJ016343 CquiOBP63	mclassic9b	2.60E-09	OBP56e	3.95E+00	OBP44a	1.43E-02
AGAP000580 AgamOBP38	-	AGAP002189 AgamOBP13	mclassic9b	3.28E-06	AGAP010409 AgamOBP22	mclassic8	8.13E-03	OBP99b	2.43E-04	OBP99c	4.20E-05
AGAP000640 AgamOBP33	-	AGAP010409 AgamOBP22	mclassic8	2.09E-10	AGAP002025 AgamOBP11	mclassic9b	2.50E-06	OBP99a	3.12E-06	OBP99c	6.15E-04
AGAP005182 AgamOBP41	-	AGAP002025 AgamOBP11	mclassic9b	6.86E-06	AGAP002025 AgamOBP11	mclassic9b	1.97E-10	OBP99a	5.57E-02	OBP56e	1.80E-04
AGAP009065 AgamOBP42	-	AGAP002025 AgamOBP11	mclassic9b	7.13E-06	AGAP002025 AgamOBP11	mclassic9b	6.62E-08	OBP99a	5.07E-05		
AGAP011647 AgamOBP30	matype1	AGAP010409 AgamOBP22	mclassic8	2.16E-10	AGAP002025 AgamOBP11	mclassic9b	1.01E-08	OBP99a	1.89E-08		
AAEL014430 AaegOBP58	-	AAEL004343 AaegOBP19	mclassic9a	6.60E-09	AAEL011730 AaegOBP81	mclassic8	6.40E-05	OBP99c	2.68E-07	OBP99b	9.63E-05
AAEL014431 AaegOBP110	-	AAEL005772 AaegOBP22	Obp99a (minus C)	1.81E-07	AAEL004342 AaegOBP18	mclassic9a	1.80E-08	OBP99b	9.50E-08		
AAEL008640 AaegOBP113	-	AAEL011730 AaegOBP81	mclassic8	2.33E-11	AAEL011730 AaegOBP81	mclassic8	1.28E-04				
AGAP010649* AgamOBP31	matype3										
AGAP010648 AgamOBP44	matype3	AGAP002025 AgamOBP11	mclassic9b	4.24E-10	AGAP002025 AgamOBP11	mclassic9b	1.90E-08	OBP99a	1.05E-05	OBP99b	1.12E-03
AGAP010650 AgamOBP45	matype3	AGAP001049 AgamOBP3	OS-E/OS-F	8.53E-12	AGAP002189 AgamOBP13	mclassic9b	2.85E-10	OBP99b	1.86E-04	OBP99a	1.57E-09
AAEL006387* AaegOBP29	matype3	AAEL002617 AaegOBP12	mclassic3a	1.08E-08	AAEL011730 AaegOBP81	mclassic8	7.44E-06	OBP56d	5.66E-05	OBP99a	3.23E-07
AAEL006398* AaegOBP32	matype3	AAEL002596 AaegOBP9	mclassic3a	9.89E-03	AAEL011730 AaegOBP81	mclassic8	3.38E-06	OBP56h	1.20E-02	OBP99a	5.77E-06
AAEL006396 AaegOBP31	matype3	AAEL002596 AaegOBP9	mclassic3a	1.06E-05	AAEL004342 AaegOBP18	mclassic9a	7.84E-03	OBP56d	1.56E-05	OBP99a	1.76E-04
AAEL006393 AaegOBP28	matype3	AAEL002617 AaegOBP12	mclassic3a	9.17E-06	AAEL004343 AaegOBP19	mclassic9a	1.22E-05	OBP56d	7.82E-05	OBP99a	1.77E-06
AAEL006385 AaegOBP33	matype3	AAEL002596 AaegOBP9	mclassic3a	4.22E-04	AAEL004343 AaegOBP19	mclassic9a	6.19E-06	OBP56d	1.34E-03	OBP99a	3.23E-07
CPIJ009038* CquiOBP86	matype3	CPIJ009937 CquiOBP44	mclassic8	2.77E-07	CPIJ009937 CquiOBP44	mclassic8	3.33E-07	OBP56d	4.89E-03	OBP99a	8.05E-10
CPIJ017342 CquiOBP92	matype3	CPIJ009937 CquiOBP44	mclassic8	3.45E-08	CPIJ017326 CquiOBP43	Obp99a (minus C)	4.00E-04	OBP56c	1.67E-02	OBP99b	6.60E-08
AGAP000641/644* AgamOBP37/34	matype2	AGAP010409 AgamOBP22	mclassic8	4.23E-07	AGAP002025 AgamOBP11	mclassic9b	1.00E-09	OBP19d	5.60E-04	OBP69a	1.42E-03
AGAP000643 AgamOBP36	matype2	AGAP000278 AgamOBP9	Obp99a (minus C)	2.20E-08	AGAP000278 AgamOB9	Obp99a (minus C)	2.20E-08	OBP56d	2.32E-05	OBP56d	2.23E-05

Mosquito Atypical OBP		Mosquito classic OBP closest homologues						Drosophila OBP closest homologues			
ID	Phylogenetic subgroup	N-term ¹	Phylogenetic subgroup	E-value	C-term ²	Phylogenetic subgroup	E-value	N-term ¹	E-value	C-term ²	E-value
AGAP000642 AgamOBP35	matype2	AGAP000278 AgamOBP9	Obp99a (minus C)	1.90E-08				OBP56d	2.23E-05	OBP69a	1.41E-02
AAEL001153* AaegOBP106	matype2	AAEL007003 AaegOBP80	-	1.07E-04				OBP99c	3.89E-05		
AAEL014876* AaegOBP107	matype2	AAEL011730 AaegOBP81	mclassic8	1.56E-09				OBP99c	1.30E-05		
AAEL013720* AaegOBP91	matype2	AAEL007003 AaegOBP80	-	3.06E-06				OBP44a	2.80E-05		
AAEL001153* AaegOBP106	matype2	AAEL007003 AaegOBP80	-	1.07E-04				OBP99c	3.80E-05		
AAEL001174* AaegOBP98	matype2	AAEL004343 AaegOBP19	mclassic9a	3.84E-06				OBP44a	5.06E-05		
AAEL001179 AaegOBP99	matype2	AAEL004343 AaegOBP19	mclassic9a	1.50E-08				OBP99b	3.15E-06		
AAEL001189 AaegOBP105	matype2	AAEL004342 AaegOBP18	mclassic9a	1.89E-07				OBP44a	6.99E-05		
AAEL004516 AaegOBP104	matype2	AAEL004343 AaegOBP19	mclassic9a	1.90E-05				OBP44a	6.86E-05		
AAEL000344 AaegOBP94	matype2	AAEL007003 AaegOBP80	-	3.05E-04				OBP44a	1.00E-05		
AAEL000319 AaegOBP93	matype2	AAEL002587 AaegOBP11	mclassic3b	1.13E-01				OBP44a	6.78E-03		
AAEL000350 AaegOBP95	matype2	AAEL011730 AaegOBP81	mclassic8	1.24E-04				OBP44a	1.25E-02		
AAEL000377 AaegOBP89	matype2	AAEL007003 AaegOBP80	-	3.05E-04				OBP44a	1.00E-05		
AAEL000318 AaegOBP92	matype2	AAEL007003 AaegOBP80	-	2.26E-06				OBP44a	5.00E-04		
AAEL014874 AaegOBP108	matype2	AAEL004343 AaegOBP19	mclassic9a	3.70E-06				OBP99c	7.93E-03		
AAEL009433 AaegOBP109	matype2	AAEL004343 AaegOBP19	mclassic9a	3.70E-06				OBP99c	7.93E-03		
AAEL013719 AaegOBP90	matype2	AAEL004342 AaegOBP18	mclassic9a								
CPIJ017166* CquiOBP94	matype2							OBP44a	8.43E-06		
CPIJ003865* CquiOBP87	matype2	CPIJ009937 CquiOBP44	mclassic8	1.06E-05				OBP44a	3.20E-06		
CPIJ017165* CquiOBP97	matype2	CPIJ009937 CquiOBP44	mclassic8	3.26E-09				OBP99c	5.57E-04		
CPIJ017167* CquiOBP95	matype2	CPIJ009937 CquiOBP44	mclassic8	8.51E-06				Obp44a	1.41E-03		
CPIJ017163 CquiOBP98	matype2	CPIJ009937 CquiOBP44	mclassic8	8.99E-06				OBP44a	6.08E-04		
CPIJ017164 CquiOBP96	matype2	CPIJ009937 CquiOBP44	mclassic8	4.10E-05				OBP44a	6.44E-04		
CPIJ017170 CquiOBP93	matype2	CPIJ017326 CquiOBP43	Obp99a (minus C)	2.26E-02				OBP44a	1.73E-03		
CPIJ003863 CquiOBP88	matype2	CPIJ010787 CquiOBP51	mclassic9a	1.01E-06				Obp44a	1.14E-08		
CPIJ003866 CquiOBP89	matype2	CPIJ017326 CquiOBP43	Obp99a (minus C)	1.77E-06				Obp44a	3.66E-09		

Mosquito Atypical OBP		Mosquito classic OBP closest homologues						Drosophila OBP closest homologues			
ID	Phylogenetic subgroup	N-term ¹	Phylogenetic subgroup	E-value	C-term ²	Phylogenetic subgroup	E-value	N-term ¹	E-value	C-term ²	E-value
CPIJ003867 CquiOBP90	matype2	CPIJ009937 CquiOBP44	mclassic8	7.88E-10				Obp44a	1.04E-06		
CPIJ001690 CquiOBP91	matype2	CPIJ009937 CquiOBP44	mclassic8	1.79E+08				Obp44a	2.80E-07		
CPIJ017169 CquiOBP99	matype2	CPIJ009937 CquiOBP44	mclassic8	5.73E-02				OBP44a	8.77E-03		
CPIJ017168 CquiOBP100	matype2	CPIJ012718 CquiOBP19	mclassic3b	7.60E-01							
AGAP002190* AgamOBP39	matype1	AGAP000278 AgamOBP9	Obp99a (minus C)	4.95E-10	AGAP002189 AgamOBP13	mclassic9b	9.88E-05	OBP99b	3.43E-08		
AGAP002191 AgamOBP40	matype1	AGAP002188 AgamOBP12	-	3.70E-08	AGAP002025 AgamOBP11	mclassic9b	1.07E-10			OBP99a	9.88E-06
AAEL009597* AaegOBP40	matype1	AAEL005772 AaegOBP22	Obp99a (minus C)	7.14E-11	AAEL004342 AaegOBP18	mclassic9a	5.22E-14	OBP99b	2.33E-11	OBP99a	2.12E-08
AAEL009599* AaegOBP41	matype1	AAEL005772 AaegOBP22	Obp99a (minus C)	1.09E-11	AAEL004342 AaegOBP18	mclassic9a	1.09E-11	OBP99a	5.69E-08	OBP99a	2.63E-03
CPIJ015732* CquiOBP84	matype1	CPIJ010787 CquiOBP51	mclassic9a	2.70E-10	CPIJ016343 CquiOBP63	mclassic9b	3.32E-17	OBP99b	1.01E-06	OBP99a	3.25E-09
CPIJ015733* CquiOBP85	matype1	CPIJ010787 CquiOBP51	mclassic9a	8.40E-10	CPIJ010782 CquiOBP46	mclassic9b	1.32E-03	OBP99a	8.77E-06	OBP99a	1.19E-06

3

Association of putative members to family of mosquito odorant binding proteins: scoring scheme using fuzzy functional templates and cysteine residue positions

3.1. Introduction

Biological sequence data are accumulating rapidly as a result of advanced sequencing technology and concerted genome projects more than the growth in computing efficiency (Butte 2001). The probability that a new protein can be classified as part of a sequence family is already near 50%. Encouragingly, evolutionary constraints on protein sequences are imposed by requirements of three-dimensional structure and biological function which are one of main aspects used for the classification of proteins. Generally, functional requirements are known to be more pronounced in terms of residue conservations, where an occurrence of completely conserved residues indicates specific biological function. Many examples of such occurrences have been reported in protein sequences: for example, the SER-HIS-ASP triad of serine proteases (Kraut 1977) and zinc finger motif of DNA binding proteins (Miller et al. 1985). Mutation of such residues generally renders the protein inactive. Such residues can be either spread across the entire stretch of the protein or can be observed as conserved patterns termed “functional motifs”. Such conservations have been used in annotating protein sequences by different methods reviewed in (Ouzounis et al. 2003). However, residues near the active site might play an auxillary role and are less easy to identify as part of ‘functional motifs’. Sequence conservation of functional residues is less obvious for residues that modulate the specificity of biological function. These residues change as a protein evolves to satisfy modified functional constraints, while the basic biochemical mechanism and the overall three-dimensional fold remain unaltered. In such cases, representative

residues associated with structural aspects of a protein serve as better classifiers. Cysteine, as a sulphur containing non-essential biogenic amino acid, plays critical roles in a number of metabolic processes. They are found as a part of a number of biological important proteins associated with important roles starting from the folding to maintaining the integrity of the structure to function. The most important role of cysteines are the formation of disulphide bridges involved in the folding of the proteins to form three-dimensional structures. Disulphide bonds, which are formed by sequentially far away cysteines but spatially proximate cysteines (Thornton 1981), define the rigidity of large globular proteins. These disulphide bonds are generally conserved among related proteins and (Richardson 1981; Srinivasan et al. 1990; Johnson and Overington 1993) and the connectivity patterns can be used to identify proteins of similar 3-D structure (Thangudu et al. 2008). The conservation of disulphide bond connectivity pattern enables the identification of remote homologs even when most of popular sequence search methods fail to do so. Such approaches are complicated by observations of topologically equivalent disulphide bonds in non-homologues and also by non-equivalent number of disulphide bonds in close homologues (Mas et al. 2001).

Owing to the fact that disulfide connectivity pattern formation in a protein is a directed (i.e. non-random) process (Benham and Jafri 1993), it can be used to obtain a structural classification of proteins. A large variety of connectivity patterns are found in disulphide-containing proteins (Benham and Jafri 1993; Harrison and Sternberg 1994). In proteins with low sequence similarity, identical connectivity patterns can indicate high structural homology. Proteins that share a disulfide bonding pattern usually belong to the same structurally derived family. Therefore, disulfide connectivity patterns provide a rapid and simple method for structural characterization of protein sequences and for examining structural properties, such as protein topologies (Benham and Jafri 1993), entropic effect of cross-linkage (Harrison and Sternberg 1994), structural superimposition of proteins by means of their disulfide bridge topology (Mas et al. 2001) and taxonomy of small disulfide-rich protein folds (Harrison and Sternberg 1994). In addition, methods that classify proteins based on their connectivity patterns have also been established. (Lenffer et al. 2004). A systematic method for the classification of disulphide-rich proteins based on cysteine conservation is thus worth undertaking. Previous attempts on cysteine based classification of proteins include approaches based on cysteine pairing (Lenffer et al. 2004), identification of odorant binding proteins based on cysteine motifs (Zhou et al. 2004), conotoxin superfamily classification using pseudo amino acid composition and multi class support vector machines (Mondal et al. 2006) and classification of peroxiredoxins using regular expressions (Chon JK 2005).

An algorithm has been devised that can efficiently identify and also classify a new protein as an odorant binding protein belonging to a particular class by capturing specific information in terms of 1) Functional residue conservation 2) cysteine conservation and disulphide connectivities. The functional residue-based scoring scheme was based on the assessment of the conservation of residues on functionally important sites in terms of sequence and a distance based scheme in terms of structure. The functionally important sites were determined by the mapping of ligand binding residues on the structural alignment of the available structural members. The test sequences were aligned to the structural alignment and scores were assigned based on the residue conservation at these functional sites. The scoring of the distance-based scheme was based on a distance criterion between the residues at these positions. The distance criteria were established by observing the distances between the residues in the functional sites, including the ‘fuzziness’ i.e. the variation in distances, among the crystal structure. The scores were calculated by a fit criterion of the distances in the models of the unknown sequences. For the cysteine-based scheme, a training “disulphide profile” of aligned sequences (Thangudu et al. 2008) has been employed of the various classes. The query sequences are aligned with these disulphide profiles followed by assigning a score based on the conservation of the cysteines in the query and further classifying them based on a composite classification scheme. This classification protocol was also implemented on the conotoxin family of proteins to extend the use of this method for the classification of disulphide-rich protein families at the subfamily level.

3.2. Methodology

3.2.1. Datasets

Seven structural entries of OBPs (PDB ID: 1dqe, 2wcj, 2gte, 2erb, 3k1e, 3bfh, 1ow4), available then, were used for the construction of the structural alignment. The dataset used in this analysis comprises of 116 conotoxin sequences (Mondal et al. 2006) and 284 odorant binding proteins from the mosquito genomes described previously in Chapter 2. The conotoxins are classified into seven classes. The odorant binding proteins are classified into three major classes *Classic*, *Plus C* and *Atypical*; the *Atypical* are further divided into 4 subtypes (*matype 1 - 4*). Representative sequences were chosen from the different classes for the construction of the training profile and the other sequences were used in the test set (Table 3.1).

3.2.2. Construction of Profiles

A structural alignment constructed using COMPARER (Sali and L.Blundell 1990) was used as a profile for the functional residue-based scoring scheme (Figure 3.1). For the cysteine-based scoring scheme, representative sequences from each class, which have conserved cysteines at all the positions under consideration, were aligned separately using ClustalW (Thompson et al. 1997). This alignment of representative sequences was used as a training profile for the classification of query OBPs. The number of sequences in the training profile and the number of cysteine positions under consideration vary for the different classes of the protein. Thus, a number of training profiles equal in number to the number of classes was generated.

3.2.3. Construction of fuzzy functional template

For the functional residue-based scoring scheme based on functional residues, a fuzzy functional template was constructed. Ligand binding residues for each of the ligand bound forms of each of the structural entries mentioned above were identified using LIGPLOT. These residues were mapped on the structural alignment (Figure 3.1). 12 residue positions were considered as functionally important positions as marked in Figure 3.1. C^β - C^β distances between residues at these positions for each of the structural entries were calculated and averaged. The upper and lower limit for the distances were set to ± 2 SD (Standard deviation) from the average distance and represented in the form of a matrix (Figure 3.2). This logic of inscribing distance variation amongst functional important residues is as adopted by Skoknick's group earlier (Fetrow and Skolnick 1998).

3.2.4. Scoring of query sequences

3.2.4.1. Functional residue based scoring scheme

Different scoring functions were defined for scoring the conservation of residues in the functional positions based on their occurrence, probability of occurrence and by consulting Dayhoff matrix.

MAJORITY BASED SCHEME:

In this, a score of 1 is given to a position in the query sequence if it has the amino acid which occurs in majority of times at that position in the structural alignment (from known observations) and finally these scores are averaged for all the 12 positions.

PROBABILITY BASED SCHEME:

A score is given to each amino acid at a position in the query sequence equal in magnitude to its probability of occurring at that position. In one scheme (OLD_PROB), the scores are finally averaged for all the 12 positions, and in the second scheme (NEW_PROB), the sum of scores is divided by the sum of the maximum probabilities of occurrence each position.

DAYHOFF MATRIX BASED SCHEME:

For each position in the query sequence, the score is calculated as the product of probability of each amino acid occurring at that position in the template and the Dayhoff Matrix score for the amino acid substitution from that AA to the residue present in the query. Finally, the scores are averaged for all the 12 positions. However, this matrix of amino acid exchanges are recorded and normalized as observed for large numbers of unrelated protein families and are also not position-specific in nature.

Given a query string Q with amino acid Q_i at functional position i , where $0 \leq i \leq p$ and a training profile T which is an alignment with i functional positions.

The scores according to the different schemes are defined as follows:

Majority based score

$$\frac{\sum_{i=1}^p \text{IsEqual}(P_i(Q_i), m_i)}{p} \quad (\text{IsEqual}(P_i(Q_i), m_i) = 1 \text{ if } P_i(Q_i) = m_i \text{ otherwise } 0)$$

Old Probability based score

$$\frac{\sum_{i=1}^p P_i(Q_i)}{p}$$

New probability score

$$\frac{\sum_{i=1}^p P_i(Q_i)}{\sum_{i=1}^p m_i}$$

$$\frac{\sum_{i=1}^p \sum_{j=1}^n M(T_{ij}, Q_i)}{p}$$

where:-

p = number of functional positions under consideration

n = number of sequences in the training profile (structure alignment)

T_{ij} = amino acid at position i in the sequence j of the training profile

Q_i = amino acid at position i of the query sequence

m_i = maximum probability of occurrence of any amino acid at position i

$M(A,B)$ = entry in substitution matrix for amino acid A being substituted by B

$P_i(A)$ = Probability of amino acid A occurring at position i in the training profile

3.2.4.2. Functional Residue Distance-Based Scoring scheme

C^β-C^β distances of the residues at the functional positions were calculated in the models (as would be described in Chapter 4) of the query sequences. The distances in the fuzzy functional template residue pairs with SD < 2 were considered for the final scoring scheme. The query sequences were aligned to the structure alignment profile and the distances between residues corresponding to the functional position were calculated in their respective models. If the distance of the residue pairs fall within the upper and lower limits assigned for those residue pairs in the functional template a score of 1 was awarded (else score is 0) and averaged for the 12 functional positions.

3.2.4.3. Cysteine-based Scoring scheme

Each query sequence was aligned separately with each of the training profiles using sequence to profile alignment method using ClustalW (Thompson et al. 1997) and checked for the conservation of cysteines. If a cysteine is found at a position, a score of '1' was given; otherwise zero. In this study, a cysteine in the query is assumed to be 'strictly conserved' if it aligns perfectly with the cysteine position in the training profile. However, according to the 'relaxed criterion', an arbitrary shift of two residues on either side of the cysteine positions in the training profile is allowed for uncertainties in the sequence alignment. In addition to the scores for cysteine conservation, an extra score of '1' is added for the conservation each cysteine pair involved in disulphide bond formation. Such position-scores are normalized for all the positions within that

class and an average score is obtained for each class for each query sequence (Figure 3.3). Thus score of a query with the training profile of each class is a measure of its likelihood of belonging to that class.

3.2.5. Composite Classification Scheme

A composite classification scheme was devised for the classification of OBPs and conotoxins based on the scores for each class, the length of the query and the distance between the cysteines involved in disulphide formation (loop spacing) (Figures 5 and 6). Thus if it is an 'N'-class problem, then for each query, there will be 'N' score parameters (one for each class), a length parameter and a variable number of loop spacing (depending upon the classes). The loop spacing (number of amino acids along the sequence between the two cysteines involved in disulphide bonding) parameter would be extremely useful to distinguish between classes with the same cysteine motif but different disulphide connectivity patterns; since it is expected that the loop spacing is more or less conserved throughout the members of a family even if other inter-cysteine distances are not.

3.2.6. Re-substitution test of the cysteine based classification scheme

The re-substitution test is one of the important methods of evaluating predictive accuracy. In this test, the training set used to generate the classifier is itself used to test the classification model. In other words, the test set is the same as the training set. The re-substitution test is extremely important because it reflects the self-consistency of an identification scheme, most importantly the algorithm.

3.3. Results

3.3.1. Functional Sites and Fuzzy Functional Template

The ligand binding residues from the bound complexes of the available PDB entries were mapped to the structural alignment generated by COMPARE (Sali and L. Blundell 1990). The positions of the alignment which had ligand binding entries in at least 4 of the 7 PDB entries were considered as functional residue positions. 12 such positions were considered as components of the functional template (Figure 3.1). The C^β-C^β distance between these 12 residues were calculated and averaged in the form of a matrix called the fuzzy functional template. The distance limits were set by (Average \pm 2 Standard Deviation). It was seen that the distances between the residues pairs

were quite variable. The distances in the matrix which had less than 2 SD were considered for the calculation of the scores. 12 such distances were identified involving 12 residue pairs in the matrix (Figure 3.2). These distances were used for the scoring function.

3.3.2. Sequence-Based Scoring scheme

3.3.2.1. *NEWPROB's scoring scheme with the addition of homologs achieves the best range and correlation*

The scores were based on the occurrence, probability of occurrence and Dayhoff matrix as described in the Materials and Methods. Different training datasets were analyzed which include 1) 7-member training set which is the initial structure alignment 2) 25-member dataset where the 7-member dataset was populated (to include evolutionary data) with one additional close homologue from each of the Mosquito genome to every member in the 7-member dataset. 3) 5-member dataset where the two mosquito crystal structures 2erb and 3k1e were removed to avoid potential bias in scoring the models (as would be described in Chapter 4 since these two structures served as templates for modeling) and 4) 18-member dataset from which the two mosquito crystal structures 2erb and 3k1e and their homologues were excluded. The range of scores for each of the method on every training set were analyzed and it was observed that the NEW probability score achieved the best range followed by the majority-based scores (Table 3.2a) and they also achieved the best correlation compared to other two methods (Table 3.2b). It was also observed that addition of homologues to the initial dataset significantly improves the range and correlation.

3.3.2.2. *All the 12 positions in the scoring scheme are equivalent in importance.*

It was important to analyze if certain functional site positions contributed more to the scores in order to provide different weights on the positions. This was done by jack-knifing each of the 12 individual positions and recalculating the scores for the initial 7-member dataset. The Pearson correlation co-efficient between the scores were calculated after removing each of the 12 residue positions (Table 3.3) and it was observed that the removal of any one position from the scoring scheme does not significantly alter the scores.

3.3.2.3. *The scores are independent of the % identity of the query sequence with the template*

Since the scoring scheme is based on the probability of occurrence of an amino acid it was required to ensure the effect of sequence identity on the scores. A histogram of the number of sequences *versus* the % identity of the sequence with the closest structural template in the dataset

was plotted (Figure 3.6). The distribution of the graph indicated that the scores are indeed independent of the sequence identity. A histogram of the no of sequences versus the % identity of the query sequence with the template was plotted and the consistently high scoring and low scoring sequences were marked on it. It was observed that the distribution of the low scoring queries and high scoring queries was independent of the sequence identity (Figure 3.10).

3.3.2.4. Structure-based scoring scheme

The structure based scoring scheme shows a good range of scores (0.3 - 1.0). However, there were low scoring sequences observed in the test cases. The scores were independent of the sequence identity to its template (Figure 3.7). But the restriction of this method is the fact that the test set were models derived from members of the training set used as templates.

3.3.3. Cysteine-based Scoring Scheme

The cysteine-based scoring scheme was found to be a more direct way for the identification of OBPs in insects and was used previously in the use of identification of OBPs. In this work, however, the scheme has been further extended to classify the OBPs in the mosquito genome. Hence, practically the algorithm not only predicts the chance of a query sequence to be a putative OBP protein, but also facilitates its classification in one of the different classes of OBPs that are described below. The OBPs are classified into four major classes i) *Classic* : which carry six conserved cysteine motif ii) *Plus C* OBPs which carry additional three conserved cysteines, iii) *Dimer* OBPs or *Atypical* OBPs which carry 2 *Classic* OBP domains and hence 12 conserved cysteines and iv) *Minus-C* OBPs which lack 2 Cys residues in comparison with *Classic* OBPs. The Dimer OBPs can be further classified as *matype1-4*; all of them hold 12 conserved cysteines except *matype2*. From the alignments used in the construction of phylogenetic trees, it was observed that the cysteine conservation patterns and spacing could play an important role in the classification of OBPs. This was analyzed by observing the cysteine conservation patterns of sequences in the test datasets when aligned to profiles constructed using a training set of each of the classes described above.

A training set for the seven different classes of OBPs (disulphide profiles) was prepared, as summarized in (Table 3.1a), by identifying representative sequences from a phylogeny of odorant binding proteins of each class. For the *Minus-C* class, the same profile for *Classic* OBPs was used but only the 1st, 3rd, 4th and 6th cysteine positions were considered. A composite classification

scheme was devised for the family of Odorant Binding Proteins incorporating the seven different scores and the length of sequence as attributes (Figure 3.4). The protocol was applied to a dataset of 284 mosquito OBP sequences (from *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*) and the class predictions were compared with the predictions of class association independently made from phylogenetic analysis. The ‘confusion matrix’ of the classes predicted by the cysteine based classification scheme *versus* the phylogeny-based classification is given in Figure 3.8a. The scheme gives an accuracy of 90.14% when compared with the phylogeny-based classification for the test set sequences. The effect of different classes to this was tested using a re-substitution test.

The re-substitution test on the training set gave accuracies of 100%, 100%, 0%, 100%, 66.66% and 100% for *Classic*, *PlusC*, *Atypical1*, *Atypical2*, *Atypical3* and *Atypical4* classes, respectively. The sequences in *Atypical1*, however, from a small group of 6 sequences and do not follow a strict conservation of cysteines as the other classes of OBPs. Hence it was difficult to classify these members by our scheme explaining the poor performance of the re-substitution test for *Atypical 1* class.

3.3.4. Application of scoring schemes on well-known superfamily of conotoxins

Since the accuracy of the classification scheme needed further convincing, the algorithm was extended to the well-known cysteine-rich superfamily of conotoxins. Conotoxins are small neurotoxic peptides found in the venom of the predatory cone snails of the genus *Conus* which act primarily by modulating the activity of specific ion channels. The mature conotoxins are characterized by the presence of multiple disulphide bonds and have been classified into seven families A, M, O, I, P, T and S again on the basis of a highly conserved N-terminal precursor sequence, disulphide connectivity and mode of action (Mondal et al. 2006). Each family is characterized by the presence of one or two characteristic patterns of disulphide crosslinks (Olivera 2002). The prominent disulphide connectivity patterns in the four major families of conotoxins are shown in Figure 3.9 and were alone used for scoring purposes.

A classification scheme was developed for conotoxins as shown in Figure 3.5, incorporating the four scores corresponding to each of the four major families. The classifier (constructed using the training set as shown in Table 3.2) was tested on a dataset of 116 conotoxin sequences obtained from (Mondal et al. 2006) and the predictions made by the scheme were compared with the known classes of the sequences in (Mondal et al. 2006). The scheme gave an accuracy of 93.1% for the test

set and the confusion matrix is presented in Figure 3.8b. The re-substitution test on the training set gave an accuracy of 100% for all the four families.

3.4. Discussion

A protein family maybe related to another protein very specifically at the subfamily level and , it can be related to more diverse proteins at the superfamily level, and it is then related to even more diverse proteins at the superfamily level. The number of common properties of the proteins at each level increases toward the subfamily level. With the increasing sequences flooding into protein databases, it is becoming highly important to characterize the existing sequence databases into groups facilitating the annotation of newly added sequences. The two main methods for the classification of proteins into families are sequence clustering and protein signatures. Methods of clustering related protein sequences by their similarity are well-established and quite rapid. However, particular functionally important residues cannot be emphasized in such alignment-based phylogenies. The concept of using protein signatures as means to facilitate protein functional classification is not new; this has been used earlier to identify and classify glutathione reductases (Fetrow and Skolnick 1998) and serine proteases. In such methods, known similarities between related protein sequences and objective methods to measure the similarities have improved the classification methods. A signature is a description of an entity and it defines the characteristics associated with only that entity. Identification of this signature from a single protein sequence is difficult; however, if a number of related sequences are aligned and evolutionary data are utilised, conserved regions can be identified. These conserved areas of a protein family, domain or functional site can be used to develop a description of the family using several different methods, including regular expressions, profiles and Hidden Markov models (HMMs). In this chapter, scoring schemes and classification have been described using functional residues as well disulphide bond patterns.

Functional residues of proteins involved in ligand binding are generally conserved through the evolution of proteins and generally considered as good classifiers of protein families and for function annotation (Innis et al. 2004). However, the efficiency may drop with protein families where significant variation of the ligand binding residues is observed among members accounting to the plasticity required to accommodate diverse ligands. Cysteine positions in protein sequences, as described above, are other evolutionarily conserved sites. They can be used as effective regular expressions in protein sequences even among distantly related proteins whose classification based

on other methods would be quite challenging. However, a sequence to sequence alignment algorithm using one representative sequence for a family would not provide sufficient accuracy accounting for the insertions and deletions observed in diverse sequences. A disulphide profile, derived from representative sequences, is more suitable for compensating the occurrences of insertions and deletions. Thus, combining the aspects of regular expressions and profile-based scoring schemes could significantly improve the quality of predictions.

The algorithm used above both based on functional residue-based identification and cysteine-based classification scheme seem to serve as good factors for identifying and classifying odorant binding proteins. Both the algorithms are based on profile and regular expression based scoring schemes which improve the identification and classification of distantly related protein families - in this case, the odorant binding proteins. The functional residue-based scoring scheme, both based on sequence and structure, exhibit a good range of scores independent of the overall sequence identity which highlights the importance of examining the conservation of functional residues. Thus, designing functional residue-based scoring schemes based on individual functional templates at family and superfamily level could serve as a better annotating protocol for newly realized sequences. The cysteine-based scoring scheme not only helps in identifying OBPs, but also aids in their classification at the subfamily level with reliable accuracy. The algorithm was also applied to yet another cysteine-rich family, where similar accuracy was observed which ensures the application of the protocol to other families. However, the necessity to build a family-specific composite classification is required.

3.5. Conclusion

Evolutionarily constricted functional and structural entities/signatures combined with family specific profile-based scoring improve the annotation and quality and can also be further extended to a subfamily level classification. The above described algorithms work efficiently for the annotation and classification of new odorant binding proteins which are indeed diverse family of proteins posing a lot of challenges on regular identification and classification algorithms. This could be extended to other diverse family of proteins. However, an in-depth analysis of every superfamily for family specific signatures and the construction of composite classification scheme at the subfamily level is required.

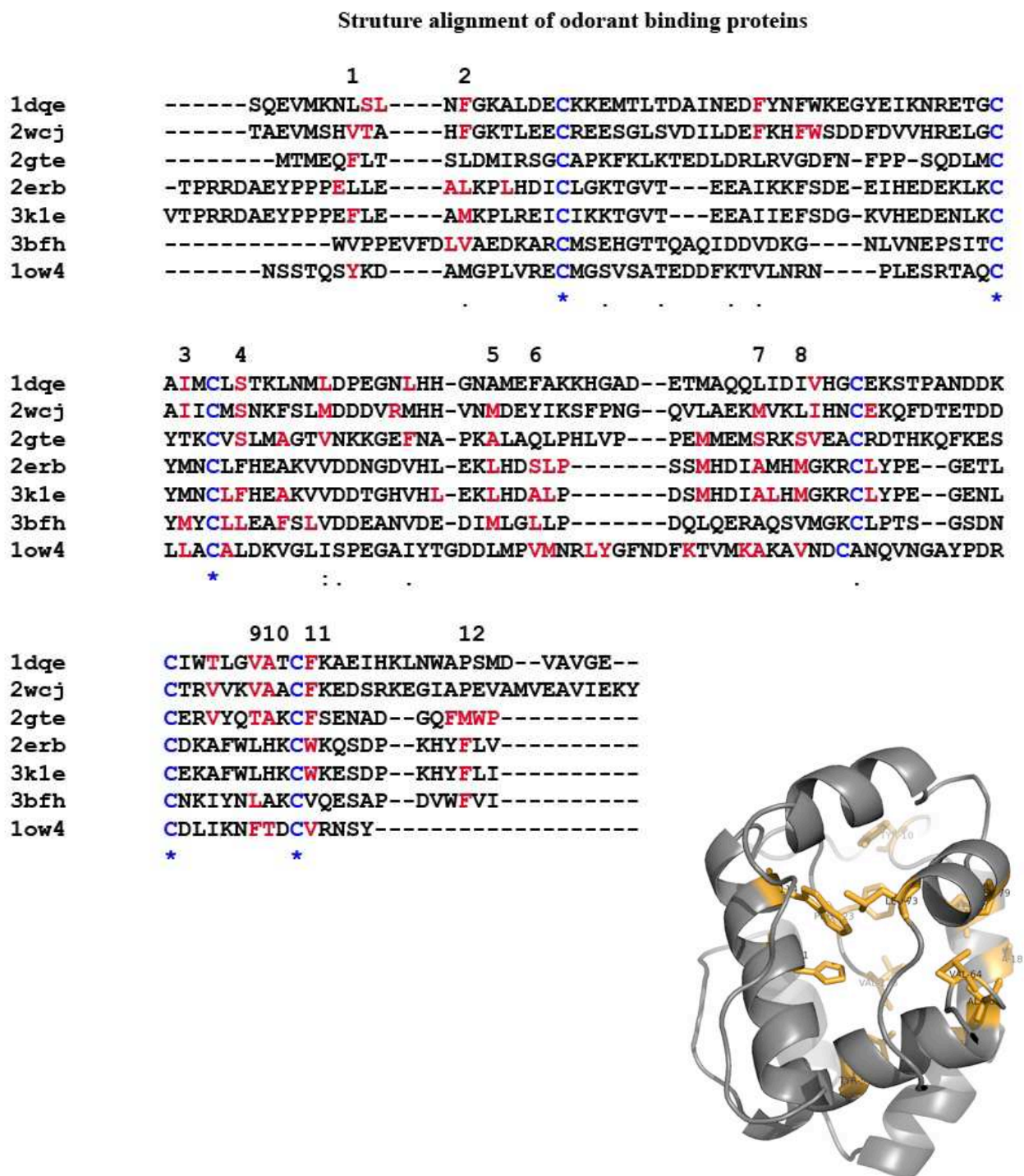


Figure 3.1. Alignment of available structures of odorant binding proteins using COMPARE. The conserved cysteines are colored in blue and functional residue is colored in red and the 12 positions used as functional sites for the scoring scheme are labelled respectively from 1 - 12 above the alignment and representation of the distribution of the functional residues on one example structure in this case 2erb.

	6	10	48	52	69	72	85	88	109	110	113	122
6	0											
10	3.61- 13.27	0										
48	13.4- 19.88	7.25- 23.77	0									
52	11.8- 19.07	6.84- 21.82	4.2- 13.07	0								
69	8.43- 20.81	8.41- 24.37	11- 19.13	5.7- 14.49	0							
71	6.55- 15.49	6.8- 10.41	11- 17.04	9- 10.83	3.5- 10.85	0						
85	8.66- 17.62	12.3- 19.2	13- 25.44	9.7- 22.68	4.8- 16.64	6.88- 18.3	0					
88	12.2- 22.42	12.5- 25.26	15- 20.37	14- 16.03	7.8- 13.48	12.9- 16.53	1.5- 14.47	0				
109	14- 24.37	11.3- 26.84	6.4- 18.3	6.6- 14.24	9.1- 12.57	13.9- 17.69	9.1- 22.86	7.4- 14.85	0			
110	11.7- 22.59	9.23- 26.59	2.4- 14.02	8- 17.27	11- 14.43	10.9- 19.74	10- 22.82	8.58- 17.81	1.92- 12.71	0		
113	12- 21.81	11.3- 23.82	10- 18.69	11- 14.35	9.6- 13.59	13.5- 16.82	4.7- 16.89	4.87- 5.7	3.83- 11.93	3.05- 15.98	0	
122	3.66- 19.71	8.6- 19.23	12- 22.18	12- 18.34	13- 19.95	13.3- 19.32	11- 13.05	10.8- 14.5	11.91- 22.91	8.54- 23.51	9.29- 12.79	0

Figure 3.2. Fuzzy functional template investigated to score the dissimilarity between OBPs.

The matrix represent the distance criteria threshold between the 12 functional sites averaged from the available structural members. The distances between pairs which have an SD<2 are colored yellow.

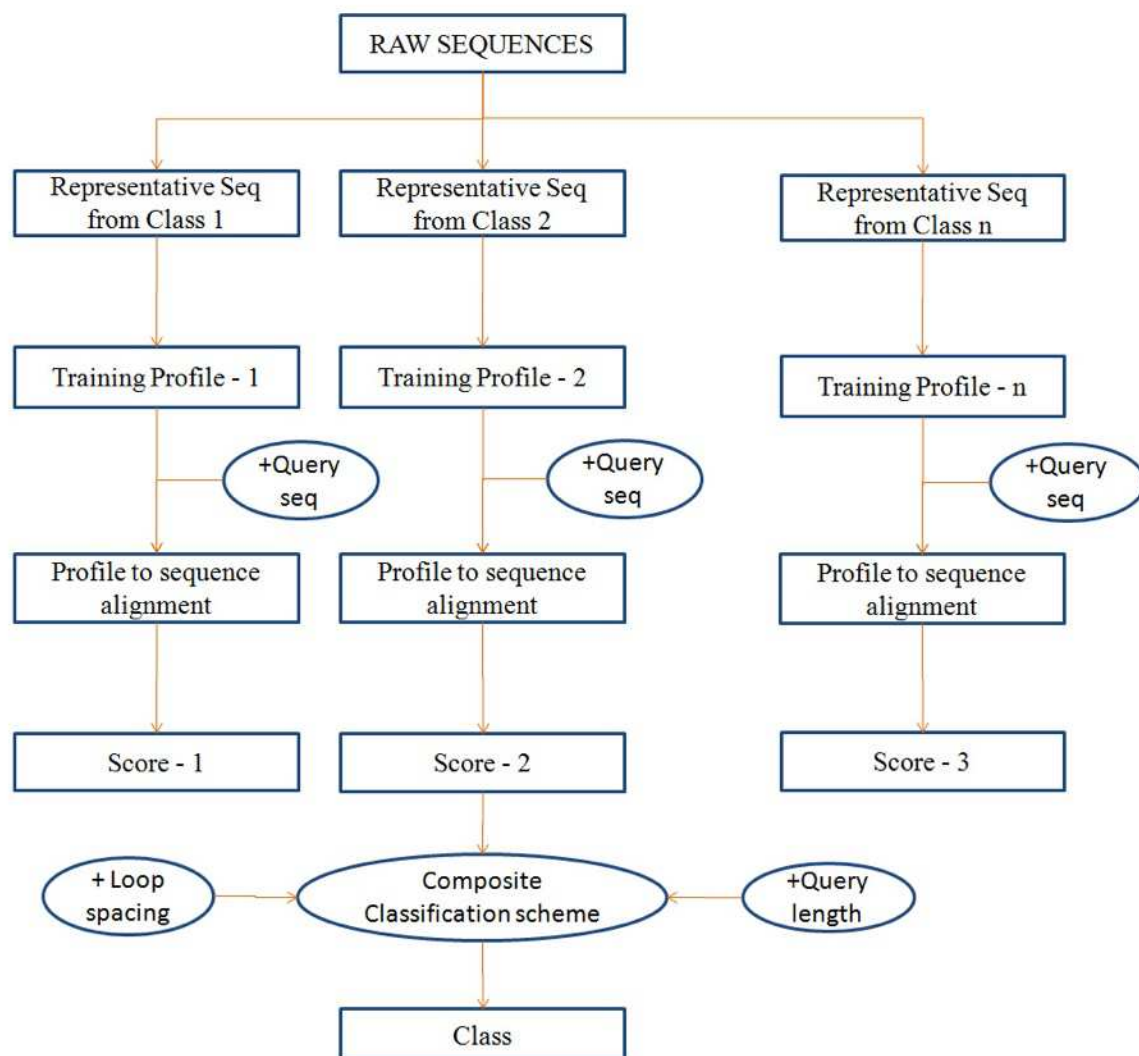


Figure 3.3. Schematic representation of the investigated cysteine based scoring scheme.

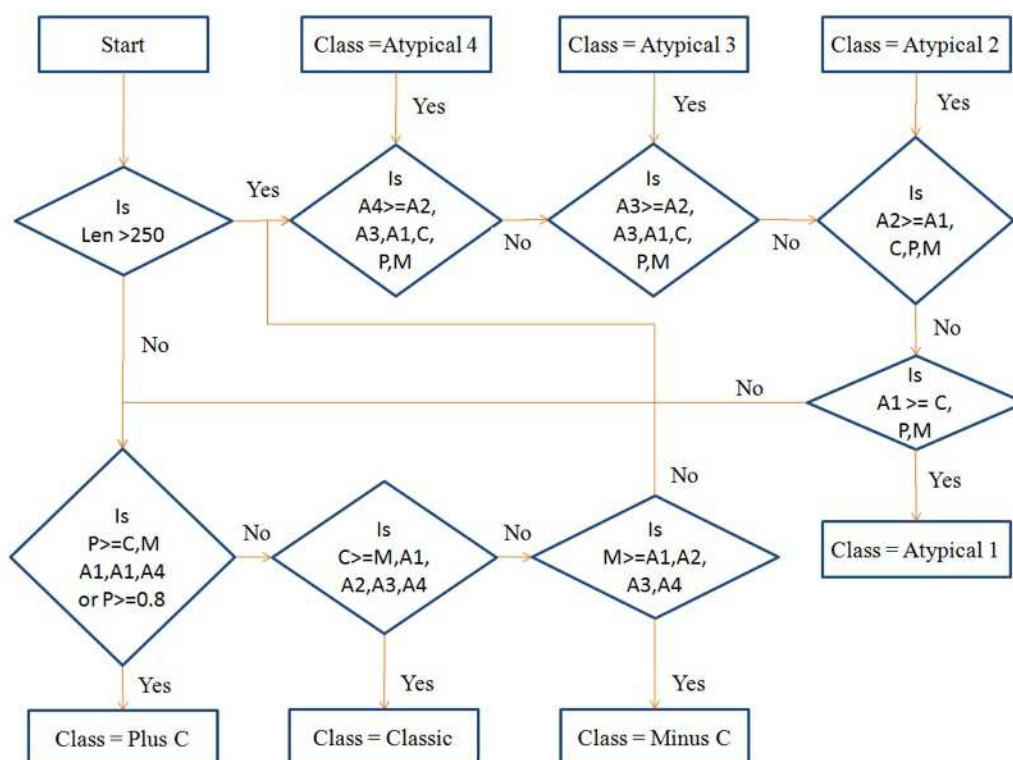


Figure 3.4. Flowchart of the logistics used in the composite classification scheme of OBPs.

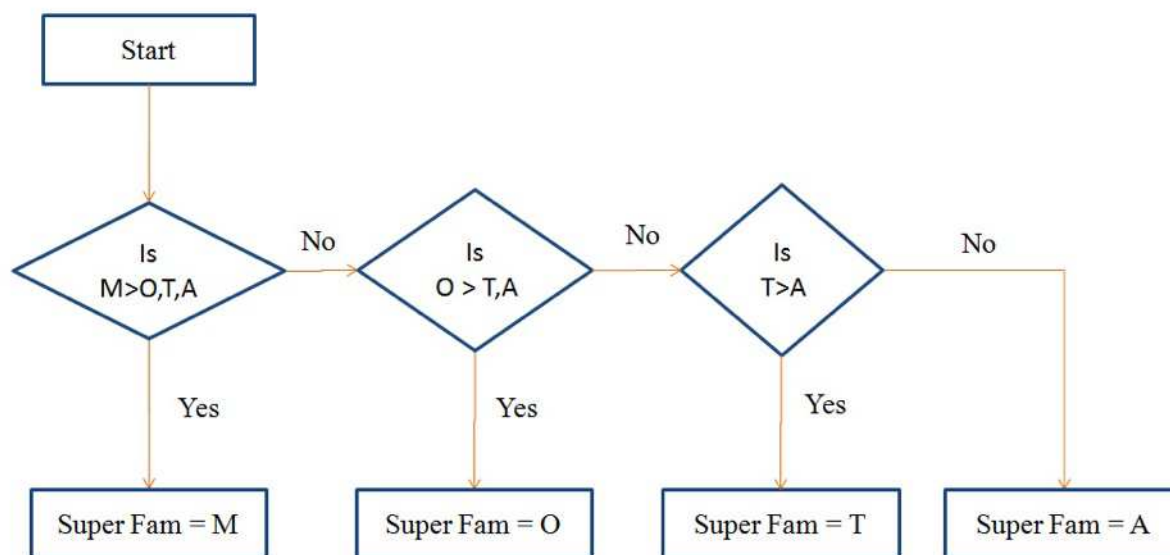


Figure 3.5. Flowchart of the logistics used in the composite classification scheme of the conotoxin family.

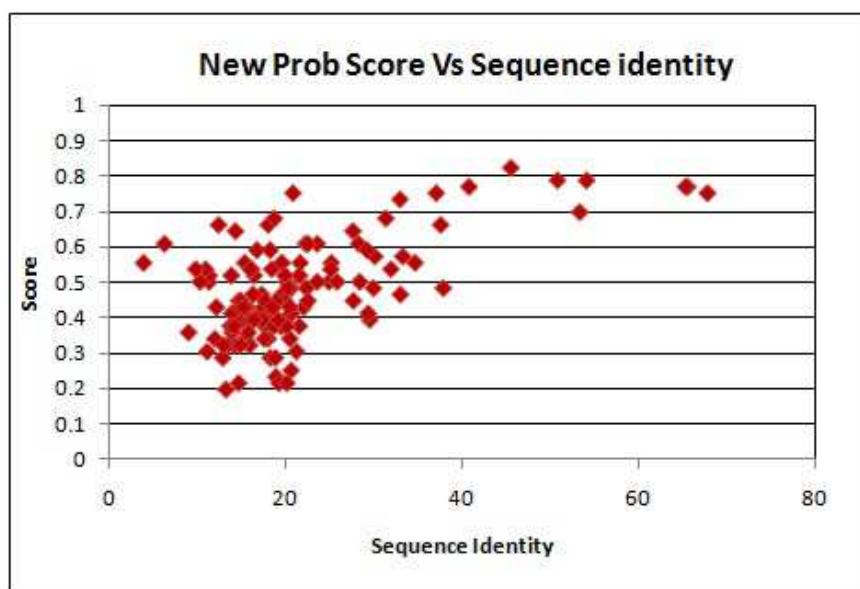


Figure 3.6. Effect of sequence identity on sequence based scoring scheme.

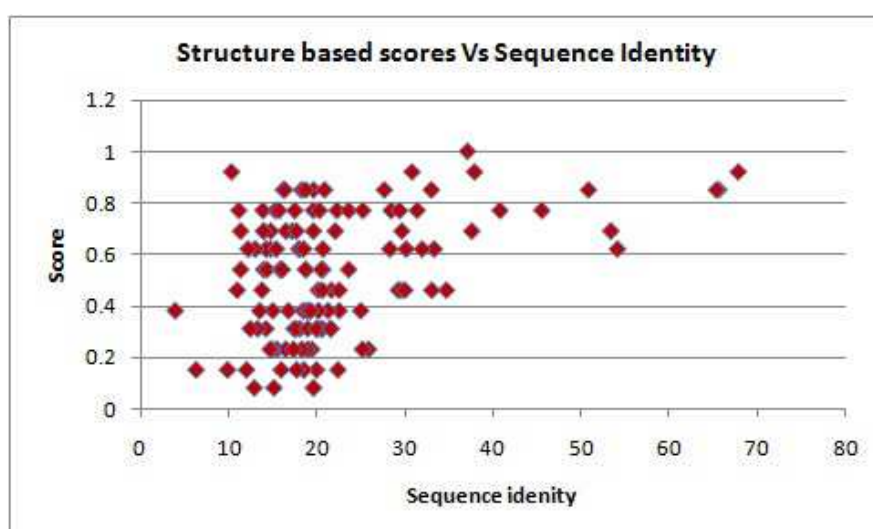


Figure 3.7. Effect of sequence identity on structure based scoring scheme.

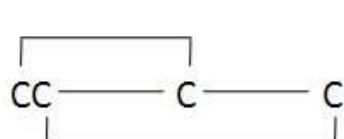
(a)

	C	P	M	A1	A2	A3	A4
C	97	1	3	0	1	0	2
P	3	45	0	0	1	0	0
M	2	0	15	0	0	0	0
A1	0	0	0	0	0	0	0
A2	0	1	2	0	21	2	0
A3	0	0	0	0	1	0	0
A4	0	0	4	1	3	3	25

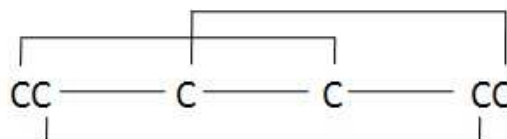
(b)

	A	M	O	T
A	18	1	0	3
M	0	7	0	0
O	0	1	53	1
T	2	0	0	9

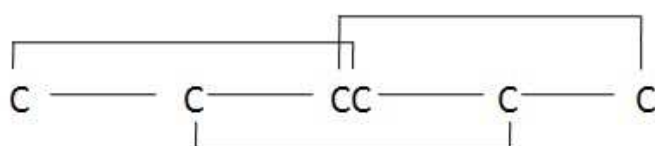
Figure 3.8. Results of the classification schemes. (a) Confusion matrix between the phylogeny based classification of odorant binding proteins and the cysteine scoring based classification scheme. (b) Confusion matrix between the classification of conotoxins and the cysteine scoring based classification scheme.



(a) Superfamily A



(b) Superfamily M



(c) Superfamily O



(d) Superfamily T

Figure 3.9. Cysteine connectivity patterns in the four major superfamilies of conotoxins. Shown are superfamily A (a), superfamily M (b), superfamily O (c) and superfamily T (d).

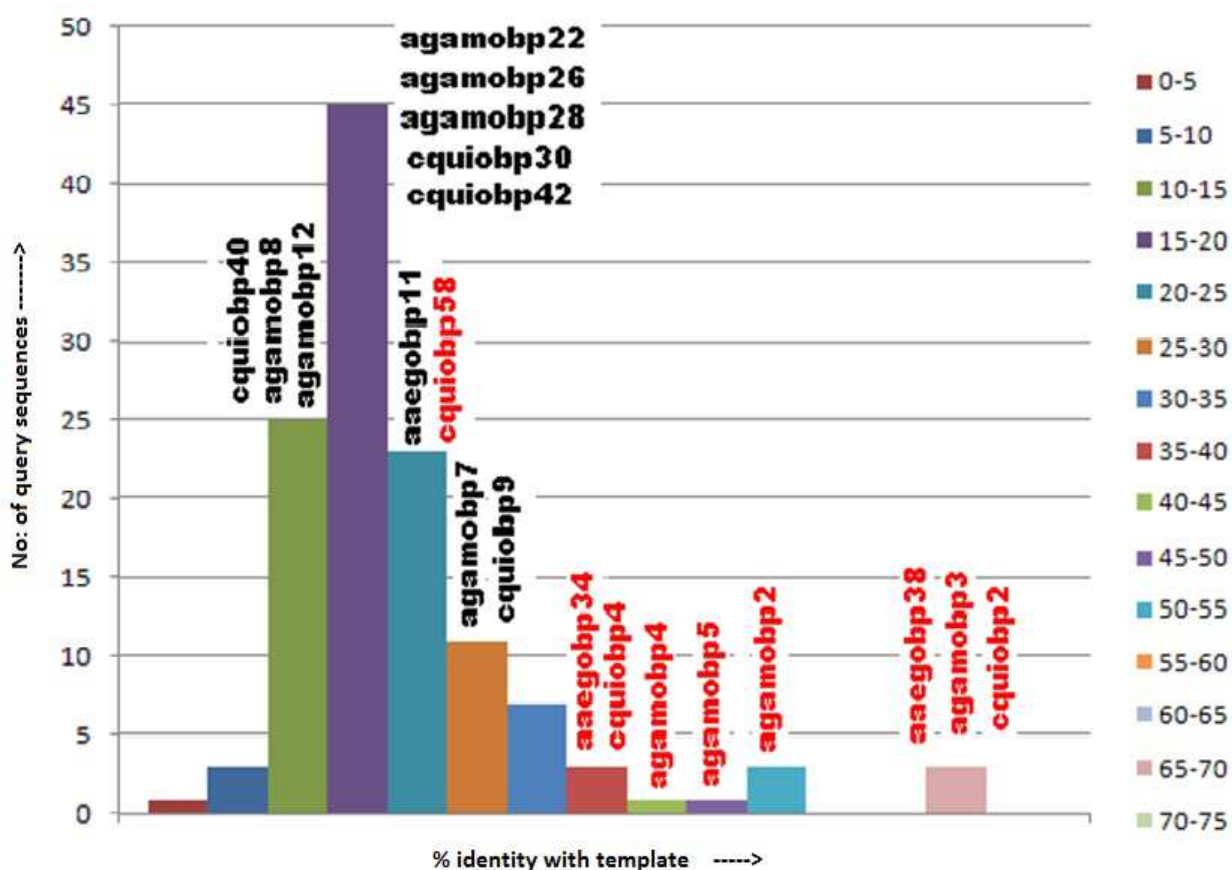


Figure 3.10. Histogram of the number of sequences versus the % identity of the query sequence with the template. The sequences labeled in red are high scoring while those labeled in black are low scoring.

Table 3.1. Datasets used as training and test sets to build and assess scorings schemes for the identification of OBPs. (a) Shown is the OBP family dataset representing the number of representative sequences used in constructing the profile (training dataset) and test set in the different classes respectively.

Protein Subfamily	Training Dataset	Test Dataset
Classic	18	104
Plus C	9	49
Minus C	18 (Classic OBPs)	17
Atypical 1	6	0
Atypical 2	6	26
Atypical 3	6	4
Atypical 4	6	33

(b) The conotoxin family dataset representing number of representative sequences used in constructing the profile (training dataset) and test set in the different classes respectively.

Protein Subfamily	Training Dataset	Test Dataset
Class A	6	19
Class M	6	7
Class O	6	55
Class T	6	11

4

Comparative modeling of classic odorant binding proteins from the mosquito genomes

4.1. Introduction

Knowledge of the native structure of a protein could provide the molecular basis for determining its function. The structure of classic odorant binding proteins (OBP) that have been deciphered so far based on crystallographic and NMR studies show that they are mainly folded into alpha helices packed compactly due to the presence of three disulphide bonds formed between the conserved cysteines in the family as described previously in Chapter 2.

The first odorant binding protein structure described was the pheromone binding protein from *Bombyx mori* both in its crystallized form, by X-ray diffraction spectroscopy (Sandler et al. 2000) and in solution using nuclear magnetic resonance (NMR) techniques (Damberger et al. 2000; Horst et al. 2001). It is represented as a rough conical structure of six helices, four ($\alpha 1$, $\alpha 4$, $\alpha 5$, $\alpha 6$) of which converge to form the hydrophobic binding pocket capped by helix $\alpha 3$. The disulphide bonds stabilize the position of helix $\alpha 3$ by attaching it to the flanking helices ($\alpha 5$ and $\alpha 6$) and the other disulphide bond bridges between $\alpha 5$ and $\alpha 6$ resulting in a rigid compact structure (Sandler et al. 2000). The compact structure however requires a conformational change in order to allow ligand binding and this was described by the change in the conformational state of the C-terminal end of the protein at different pH conditions. In acidic conditions the C-terminus of the protein folds into an α -helical domain and enters the bombykol binding site assisting the release of bombykol from the binding pocket.

However in the subsequent structural report of a pheromone binding protein from cockroach *L. maderae* described by (Lartigue et al. 2004) the requirement for an active mechanism for releasing the ligand as the pheromonal blend of the *L. maderae* is not emphasized. The structure is reported to mostly be composed of hydrophilic compounds unlike the moth pheromones. The PBP of *L. Maderae* also lacks the C-terminal segment conserved in the lepidopteran OBPs. Since this

PBP is 19 residues shorter than *B.mori* PBP (BmorPBP1) and functional, it appears to rule out possibility of a mechanism where the seventh helix could be implicated in pushing its ligand out of the binding cavity. In the meantime the structure of PBP of the giant moth *A. polyphemus* became available (Zubkov et al. 2005) and this is closely related to the BmorPBP1. In this report, the authors propose a pH-induced structural change, attributed to the protonation of His69, His70 and His95 in the binding pocket which could cause a reorientation of α -helices 1, 3 and 4, thus providing the driving force for the release of the pheromone molecule from the cavity is described (Zubkov et al. 2005).

The structure of LUSH, another OBP identified in *Drosophila*, shows its C-terminus folded back into the core of the protein and forms a part of binding cavity (Kruse et al. 2003; Thode et al. 2008). Such a conformation is similar to that assumed by moth PBPs in acidic conditions, but occurs in this protein at neutral pH. Later it was described that LUSH directly activates the pheromone receptors without the release of the ligand (Laughlin et al. 2008). ASP1 is an odorant binding protein from honeybee and is observed to be shorter than in *B.mori* PBP but longer than in *L.mandare* PBP. Like LUSH, the structure of ASP1 shows that the C-terminus folds back into the protein core without forming an α -helix, and partially occupies the binding cavity.

The first crystal structure of an OBP, AgamOBP1 from the mosquito genome was reported by (Wogulis et al. 2006). This structure was solved at a resolution 1.5Å and observed as a crystallographic dimer. The binding site is tunnel shaped at the dimer interface. A precipitant PEG molecule was found in the binding site in this structure. The structure was found to retain a similar fold compared to six other OBP structures described previously but still showed an RMSD of 4.2Å for ApolPBP1, 2.4Å for BmorPBP1, 2.3Å from LmaPBP1, 1.6Å for LUSH and 1.7Å for Amel-ASP1. The differences were mainly observed in the loop regions. The most distinguishing feature of this protein was the C-terminal loop which makes a part of the wall of the binding pocket. The carboxylate oxygens of the C-terminal are found within hydrogen bond distance with His23 and Tyr54. The dimer interface is formed across the non-crystallographic two fold axis and primarily engages the 4th and 5th helices and the loop that is C-terminal to the fifth helix. However sparse hydrophobic side chains that are observed at the interface, the absence of a clear dimer in the case of other OBPs and a non conserved interface as with LUSH suggest that the protein is more likely to be a monomer *in vivo*.

The next structural report of a mosquito OBP was that from *Aedes aegypti* (Leite et al. 2009). This sequence shares 82% identity with the previously discussed gene product AgamOBP1. The two structures showed an RMSD of 0.29 to 0.40 (involving the two chains). The structure in

this case is clearly a crystallographic dimer and it is suggested that mosquito OBPs exist in monomer-dimer equilibrium, with isolated dimers slowly converting to monomers. This structure also shows that OBP harbours the same serendipitous ligand PEG. The differences in the structure were observed at the residues of the binding pocket which can attribute to differences in ligand specificity. Since the C-terminal region was implicated in ligand binding, a detailed comparison at this region shows that this region of the two sequences is identical to AgamOBP1. A difference is observed only at the terminal residue of Ile125 instead of Val125 with the carboxylate oxygens still at a hydrogen bonding distance to Tyr54 and His23 similar to AgamOBP1 (Leite et al. 2009).

A subsequent and a very recently deciphered structure of mosquito OBP is that of CquiOBP1 (Mao et al. 2010). This sequence shares a sequence identity of 90% and 87% with the previously identified AgamOBP1 and AaegOBP1, respectively and shares similar structural features. CquiOBP1 was found to exist in monomer-dimer equilibrium in solution. The most interesting aspect of this structure was the presence of a true ligand an oviposition pheromone (5*R*, 6*S*)-6-acetoxy-5-hexadecanolide (MOP) compared to the previous structure which housed only a PEG molecule. Nevertheless the binding pocket of the three proteins did not differ even if it was a PEG molecule in place of a true ligand in AgamOBP1 and AaegOBP1. The structure described is a non-crystallographic dimer with two molecules of MOP bound to each monomer beginning at the tunnel close to the dimer interface. However a solution structure of this protein showed a dissociation of the dimer to form monomeric structures at pH 7.0 while retaining the binding of the ligand. Similar to the other mosquito OBP structures, the C-terminus extension of CquiOBP1 folds inside the central cavity, making up part of the central cavity wall. The same hydrogen bonding triad formed by the carboxylate oxygens with Tyr 54 and His23 was observed and was speculated to undergo a pH-dependent disruption resulting in the displacement of the C-terminal from the binding pocket releasing the ligand (Mao et al. 2010).

Another set of crystal structure of the *Classic* AgamOBP22 was very recently deposited in the PDB in 2011 but the description of this structure is not published so far. These include the protein complexed with glycerol (PDB:3L4A), benzaldehyde (PDB:3L4L), cyclohexanone (PDB: 3L5G & 3QME). Another entry (PDB:3PJI) from the same protein in the unbounded open status for ligand binding is indicated to have been deposited too. However the structure from the PDB shows a slightly different fold compared to the other structures.

Very recently the crystal structure of a *Plus C* OBP (AgamOBP47) (PDBID: 3PM2) was described in *Anopheles gambiae* (Lagarde et al. 2011). Similar to the classic OBPs the structure was mostly helical; however eight helices could be observed in this structure when compared to the

classic OBPs which have only six helices. Three disulphide bonds are formed between the N-terminal and C-terminal segments of the protein, two disulphides connecting helix3 and helix7 and one disulphide bond between helix1 and a β -turn loop. The structure also retains a β -turn loop that were previously found in *Classic* OBPs (Laughlin et al. 2008; Pesenti et al. 2008). When this PlusC OBP was superposed on LUSH structure, five helices superpose well, while helix2 tends to be structurally non-equivalent in the two proteins. In addition two extra helices are seen in the case of AgamOBP47. A firm conclusion on the dimerization state of this protein is not addressed and it is assumed that they might dimerize as homodimers or heterodimers.

A total of 62 structures of OBPs and PBPs, including ligand-bound and mutant forms, are available on the Protein Data Bank from different organisms.(Damberger et al. 2000; Sandler et al. 2000; Horst et al. 2001; Lee et al. 2002; Kruse et al. 2003; Lartigue et al. 2004; Mohanty et al. 2004; Lautenschlager et al. 2005; Zubkov et al. 2005; Wogulis et al. 2006; Damberger et al. 2007; Lautenschlager et al. 2007; Pesenti et al. 2008; Thode et al. 2008; Pesenti et al. 2009; Mao et al. 2010)

However considering the diversity of these proteins and a highly dispersed ligand space it is required to study individual proteins to obtain a clearer picture towards function. The modeling of all the classic odorant binding proteins from *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus* genomes is described in this Chapter and their subsequent use in functional analysis is described later in Chapter6.

4.2. Materials and methods

Comparative protein structure modeling has been used widely in the prediction of protein structures as it results in most accurate, detailed and explicit models of unknown structure s computationally in the absence of experimental data. This maximizes their usefulness in applications such as interpretation of the existing functional data, design of ligands, and construction of mutants and chimeric proteins for testing new functional hypotheses(Johnson et al. 1994). Comparative protein modeling was used to model all the Classic OBPs from *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*. Comparative modeling is a multistep process which is described below.

4.2.1. Retrieval of target sequences

The amino acid sequences of the classic odorant binding proteins of *Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus* were obtained from the VectorBase (<http://www.vectorbase.org/>). The sequences of all the OBPs were submitted to the Signal P server (<http://www.cbs.dtu.dk/services/SignalP/>) for prediction of signal peptides. The predicted signal peptide region of the sequences was removed before the modeling process since they do not form a part of the matured protein.

4.2.2. Identification of template and alignments

An attempt was made to find a suitable template for the modeling of the target sequence using BLAST. But due to the diversity of sequence that is observed with this family of proteins this procedure was found to be not the most satisfactory approach for template search. Alternatively the template protein was searched through fold recognition methods using 3DJury metaserver, http://meta.bioinfo.pl/submit_wizard.pl which is an online consensus tool for searching homologues with known information on structure, that retain both sequence and structural similarity. The top ranking homologue which is an OBP with structural information known, was downloaded along with the suggested alignment for every sequence that was modelled. The alignments as suggested by 3D-Jury were then used for modeling. In cases where N-terminal and C-terminal overhangs were observed in the query sequence, they were pruned with respect to the template.

4.2.3. Modeling and energy minimization

A rough 3-D model was constructed for every OBP in the dataset by extracting distance and dihedral angle restraints from the template structure and alignment of the target sequence with the template with the help of MODELLER 9v1 software (Sali 1995). The rough model constructed was then solvated and subject to energy minimization applied. All protein atoms were permitted to participate in energy minimization using the steepest descent and conjugate gradients to eliminate bad contacts between protein atoms and structural water molecules, in order to construct models that satisfied all the spatial restraints possible. Computations for the energy minimization were carried out using Gromacs software (Van Der Spoel et al. 2005) and OPLS-AA forcefield parameters on all atoms (William L. Jorgensen 1996).

4.2.4. Evaluation of refined model

The refined structures of the models were further evaluated for testing its internal consistency and reliability. Backbone conformation was evaluated by the inspection of the Psi/Phi Ramachandran plot, as a part of PROCHECK (Laskowski et al. 1996) (<http://biotech.ebi.ac.uk:8400/cgi-bin/sendquery>) analysis.

4.3. Results

4.3.1. Template selection and alignment

The percentage identity of the OBP sequences ranges from 10% to 90 % with their closest known structural homologue of known structural information, as determined by a consensus fold prediction method (Figure 4.1). It was observed that more than half of the OBP sequences were found in the ‘twilight zone’. However the OBPs are believed to have very conserved fold provided by conserved disulphide bonds. These cysteine residues are therefore expected to provide safe anchors (equivalences) that can be relied upon for sequence-structure alignments. The presence of a strong structural similarity and Cys-residue equivalences should hopefully allow accurate modeling even at very low sequence identity (Thangudu et al. 2005). However the modeling of the loops and terminal segments of the protein is still challenging even when the structural core of the protein can be modeled with a considerable accuracy. Models of CquiOBP55 and AaegOBP83 were alone not constructed since the sequences were very divergent from the available templates.

4.3.2. Model accuracy

As shown in Figure 4.3, 131 models were generated with the best template chosen from a consensus fold prediction approach described above and using Modeller 9v1 (Sali 1995). The template used and their respective identities with the query sequences are presented in Table1. The models were validated in PROCHECK based on the Ramachandran plots of phi-psi angles. The percentage of allowed and disallowed regions of the models was analyzed and is presented in Table1. It was observed that very few residues have phi-psi angles in the disallowed regions and this was independent of the percentage of sequence identity between the query and the template (Figure 4.2).

4.3.3. Structure analysis of members in a subfamily

The superimposition of the models of all the members in every cluster overwhelmingly indicated that the third helix, fourth helix and the loop connecting the third and fourth helix are highly conserved in terms of spatial orientation compared to other parts of the proteins (Figure 4.4). It is also interesting to note that models belonging to a cluster, independently of the template used, were closely related to each other in structural space. The models belonging to the OSE/OSF cluster were found to be the most accurate compared to the other clusters as they hold three experimental crystal structures which act as very good templates for the other members in this cluster. The helices, loops and also the terminal regions of the protein of the modeled structures were well-defined in this cluster. Similar observations were observed for the members closely related to the *OSE/OSF* cluster belonging to *LUSH* and *PBPRP1* where the structure was quite well-defined. However the N- and C-terminal segments could not be modelled with reliable accuracy. Members of *OBP19a* and *Pbprp2/Pbprp5* clusters showed a good conservation of helix 3 and the loop connecting helix3 and helix4, even if the entire structure superposition is not very good. Members belonging to different clusters from *MClassic1* – *MClassic9* showed considerable consistency in the helical regions and the loop connecting helix 3 and 4. It was interesting to note that the members in the *Bombyx mori* *Minus C* cluster also showed a rigid superposition of helix1, helix3, helix4, helix6 and the loops connecting helix3-helix4 and helix5-helix6. The structural conservation of a loop over a large group of proteins in a family is a striking feature and such conservation in most of the cases is attributed to a functional role. It would be interesting to further investigate the role of this loop connecting helix3-helix4 and helix5-helix6 experimentally for functional implication in the OBP gene family. Only in November 2010, some time after the modeling reported in this chapter was completed (September 2010), the crystal structure of CquiOBP1 bound to 3OG at pH 8.2 was published by (Mao et al. 2010) and referred in the PDB under the identification code 3OGN. The overall deviation *rmsd* between the C-alpha atoms of crystal structure and the model for CquiOBP1 determined is 0.32 Å (Figure 4.5a). Similarly, very recently on 3rd of August 2011, a structure of AgamOBP4 (PDB:3Q8I) bound to indole at pH 6.97 was published by Davrazou et al. (2011) Our model was also in good agreement with the crystal structure with a measured *rmsd* value of 0.95 Å (Figure 4.5b).

4.4. Discussion

An immediate challenge ahead for all biologists, once the sequence information of a protein is available, is to narrow down on the function of a protein and determination of the structure of a protein stands as an essential intermediate in this procedure. The importance of computational methods was not quite valued or looked upon until a recent flood of sequence information hit the biological community. Computational methods to analyze a new sequence in terms of structure and function are now being currently highly explored and many sophisticated methods are already available for addressing this problem. Among the various structure prediction methods that are currently available, comparative modeling results in the most accurate, detailed, and explicit models of protein structure. However the accuracy of the model produced is directly proportional to the similarity of the query sequence with its corresponding template. Fortunately, a 3D model does not have to be absolutely perfect to be helpful in biology (Johnson et al. 1994). One reason is that knowing only the fold of a protein is frequently sufficient to predict its approximate biochemical function. The functional prediction of a protein is most directly determined by the shape of the binding pocket rather than its sequence alone where sequentially distant residues may not, in the binding pocket follow the same order as found in a structural space. A collection of experimentally determined complexes of proteins aligned with comparative models for the rest of the family members, will permit a comparison of ligand-binding requirements. This has been found to be very useful in the process of drug design. It is also observed that the sequences belonging to a particular cluster/subfamily within a family are more likely to be structurally similar. It is also intriguing that certain parts of the protein are highly conserved spatially in spite of high sequence divergence kindling the role of such regions in a protein family as a whole.

The ultimate validation of any protein structure model is to compare it with a subsequent experimentally derived structure. In the case of CquiOBP1 (Mao et al. 2010) and AgamOBP4 (Davrazou et al. 2011) the atomic coordinates derived from crystallographic data were made available in the PDB weeks or months after the models for these proteins were constructed. Though we are considering that comparative modeling of *Classic* OBPs might generate low resolution models due to high sequence divergence, in fact, the structural constraints imposed by the constitutive disulphide bonds do participate towards a better precision of the constructed models as it was demonstrated earlier by the group (Thangudu et al. 2005). This explains in part the good agreement between the models for CquiOBP1 and AgamOBP4 and their corresponding experimentally derived structures (3OGN and 3Q8I).

Overall, despite the diversity of the family, we consider that the accuracy of our predicted structures as good considering the inherent restraint imposed by the disulphide bridges. We believe the generated structural data is hence exploitable for further analysis namely for docking experiments as seen in the chapter 6 and will provide an obvious resource for many other important questions and hopefully, will provoke new ones.

4.5. Conclusion

Elucidating odorant binding protein function is one the central focus of the biology of insects olfaction today, and computational approaches have become more important in this challenge. Understanding the molecular function of odorant binding proteins is greatly enhanced by insights gained from their three-dimensional structures. Since experimental structures are only available for a small fraction of these OBPs, the advantage of computational methods for protein structure modeling was used in addressing this issue. Although it is not possible to model all OBPs with equivalent accuracies, the current comparative modeling of *Classic* OBPs will efficiently complement their sequence analysis and associated experimental data even though they are insufficient on their own to provide strong functional insights. These predicted structures might stand to be good starting points for further experiments. As a service to the community, a database dedicated to mosquito OBPs is being set up where these models will be freely available.

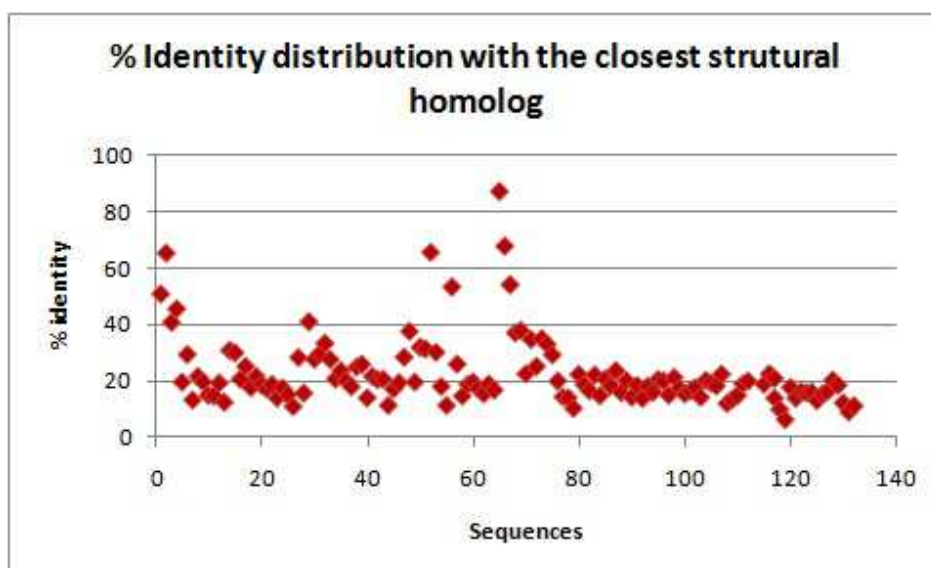


Figure 4.1. Distribution of percentage identity of the OBP sequences with their respective structure template used for modelling .

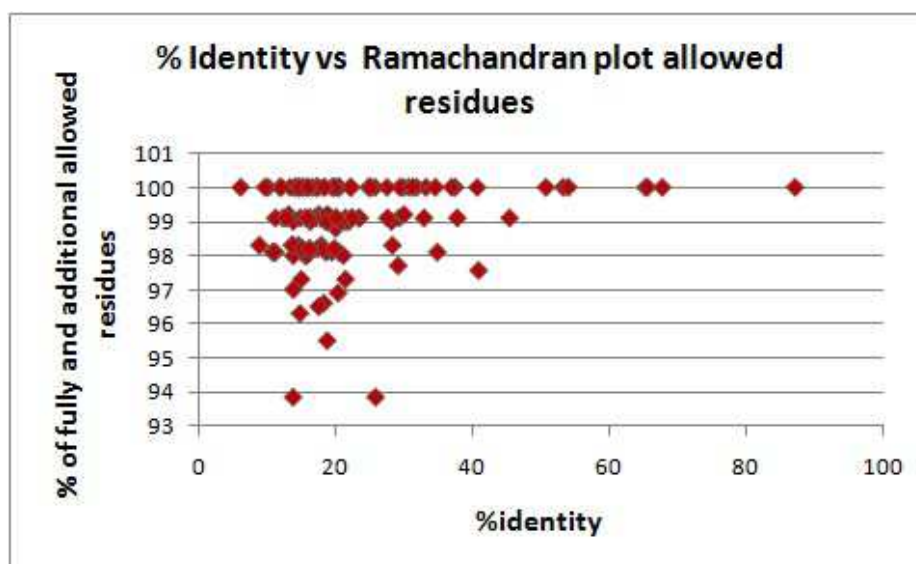


Figure 4.2. Quality assessment of modelled OBPs as a function of sequence identity. Plot of sequence identity of the query sequence and template against the sum of fully and additionally allowed phi/psi angles measured based on the analysis of Ramachandran plot for every model as a measure of the quality of the model.

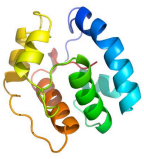
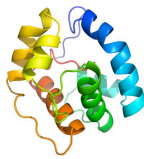
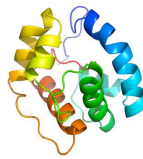
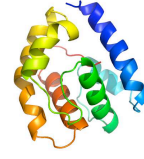
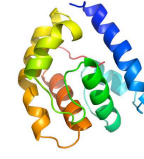
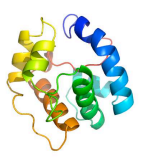
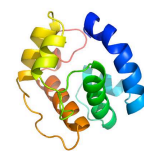
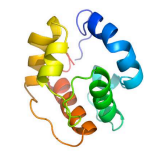


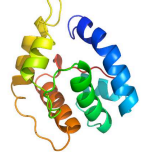
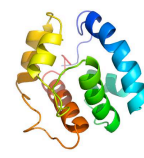
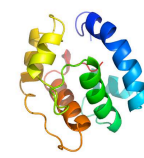
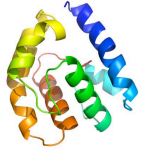
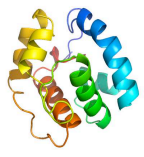
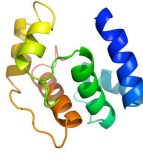
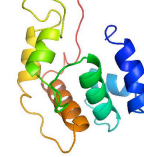
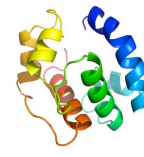
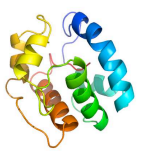
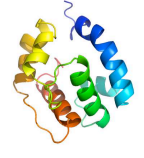
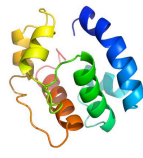
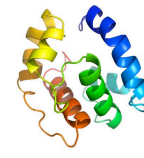
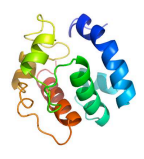
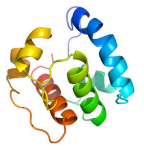
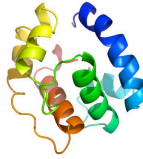
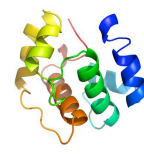
AgamOBP1	AgamOBP2	AgamOBP3	AgamOBP4	AgamOBP5
				
AgamOBP6	AgamOBP7	AgamOBP8	AgamOBP9	AgamOBP10
				
AgamOBP11	AgamOBP12	AgamOBP13	AgamOBP14	AgamOBP16
				
AgamOBP19	AgamOBP20	AgamOBP21	AgamOBP22	AgamOBP23
				
AgamOBP24	AgamOBP25	AgamOBP26	AgamOBP27	AgamOBP28
				
AgamOBP29	AgamOBP62	AgamOBP63	AgamOBP64	
				

Figure 4.3. Graphical representation from PyMOL of all the *Classic* OBP models that were constructed using MODELLER (continued on next page).

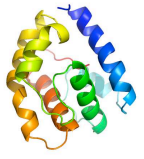
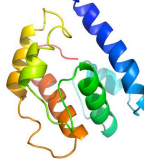
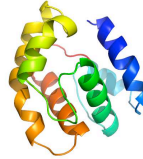
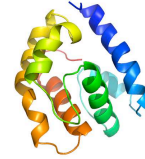
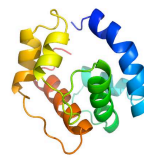
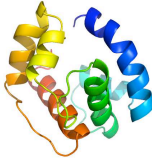
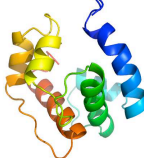
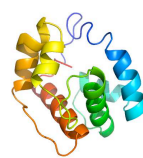
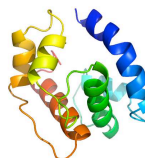
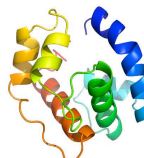
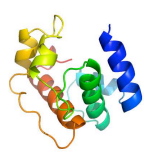

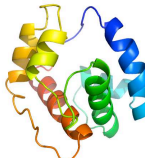
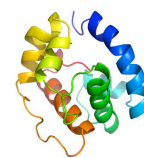
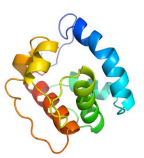
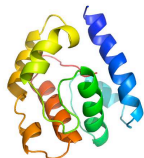
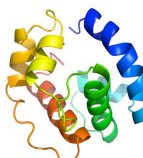

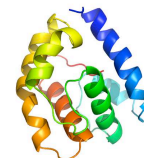
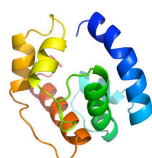
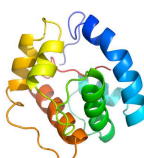

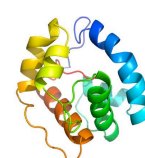
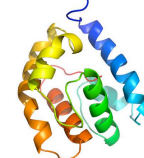
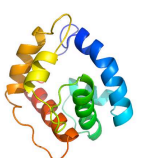

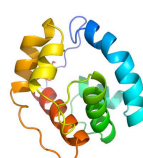
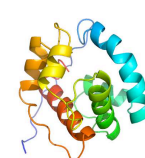

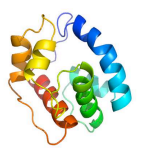
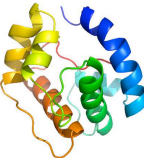
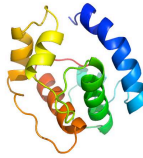

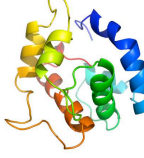
AaegOBP1	AaegOBP2	AaegOBP3	AaegOBP4	AaegOBP8
				
AaegOBP9	AaegOBP10	AaegOBP11	AaegOBP12	AaegOBP13
				
AaegOBP14	AaegOBP15	AaegOBP17	AaegOBP18	AaegOBP19
				
AaegOBP20	AaegOBP21	AaegOBP22	AaegOBP27	AaegOBP34
				
AaegOBP35	AaegOBP36	AaegOBP37	AaegOBP38	AaegOBP55
				
AaegOBP57	AaegOBP59	AaegOBP60	AaegOBP61	AaegOBP65
				
AaegOBP76	AaegOBP77	AaegOBP78	AaegOBP79	AaegOBP80
				

Figure 4.3 (contd). Graphical representation from PyMOL of all the *Classic* OBP models that were constructed using MODELLER (continued on next page).


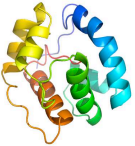
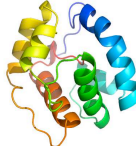


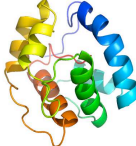
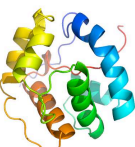
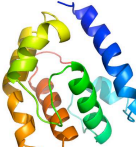
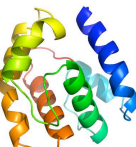
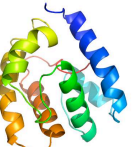


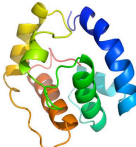

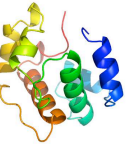
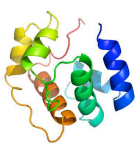
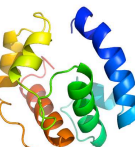
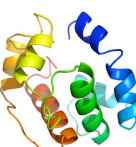
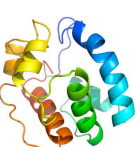
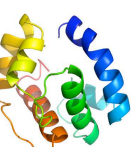
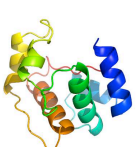
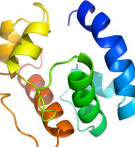
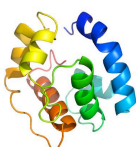
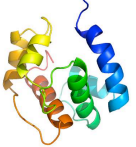



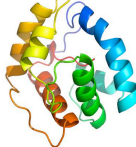
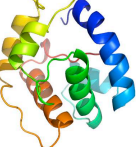
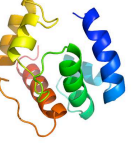

AaegOBP81				
				
CquiOBP1	CquiOBP2	CquiOBP3	CquiOBP4	CquiOBP5
				
CquiOBP7	CquiOBP8	CquiOBP9	CquiOBP10	CquiOBP11
				
CquiOBP12	CquiOBP13	CquiOBP14	CquiOBP15	CquiOBP16
				
CquiOBP17	CquiOBP18	CquiOBP19	CquiOBP20	CquiOBP21
				
CquiOBP22	CquiOBP23	CquiOBP24	CquiOBP25	CquiOBP27
				
CquiOBP28	CquiOBP29	CquiOBP30	CquiOBP31	CquiOBP32
				

Figure 4.3 (contd). Graphical representation from PyMOL of all the *Classic* OBP models that were constructed using MODELLER (continued on next page).

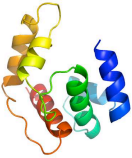
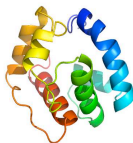
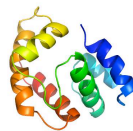




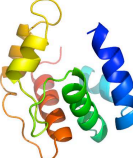

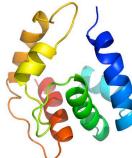
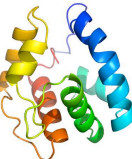

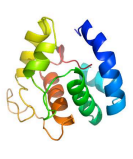
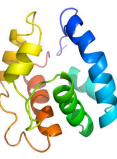
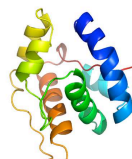








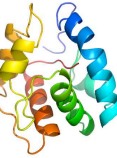






CquiOBP33	CquiOBP34	CquiOBP35	CquiOBP36	CquiOBP37
				
CquiOBP38	CquiOBP39	CquiOBP40	CquiOBP41	CquiOBP42
				
CquiOBP43	CquiOBP44	CquiOBP46	CquiOBP51	CquiOBP52
				
CquiOBP53	CquiOBP54	CquiOBP56	CquiOBP57	CquiOBP58
				
CquiOBP59	CquiOBP60	CquiOBP61	CquiOBP62	CquiOBP63
				
CquiOBP64	CquiOBP65	CquiOBP66	CquiOBP67	CquiOBP68
				

Figure 4.3 (contd). Graphical representation from PyMOL of all the *Classic* OBP models that were constructed using MODELLER (continued on next page).




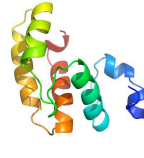


CquiOBP69	CquiOBP70	CquiOBP71	CquiOBP72	CquiOBP73
				
CquiOBP74				
				

Figure 4.3 (contd). Graphical representation from PyMOL of all the *Classic* OBP models that were constructed using MODELLER.

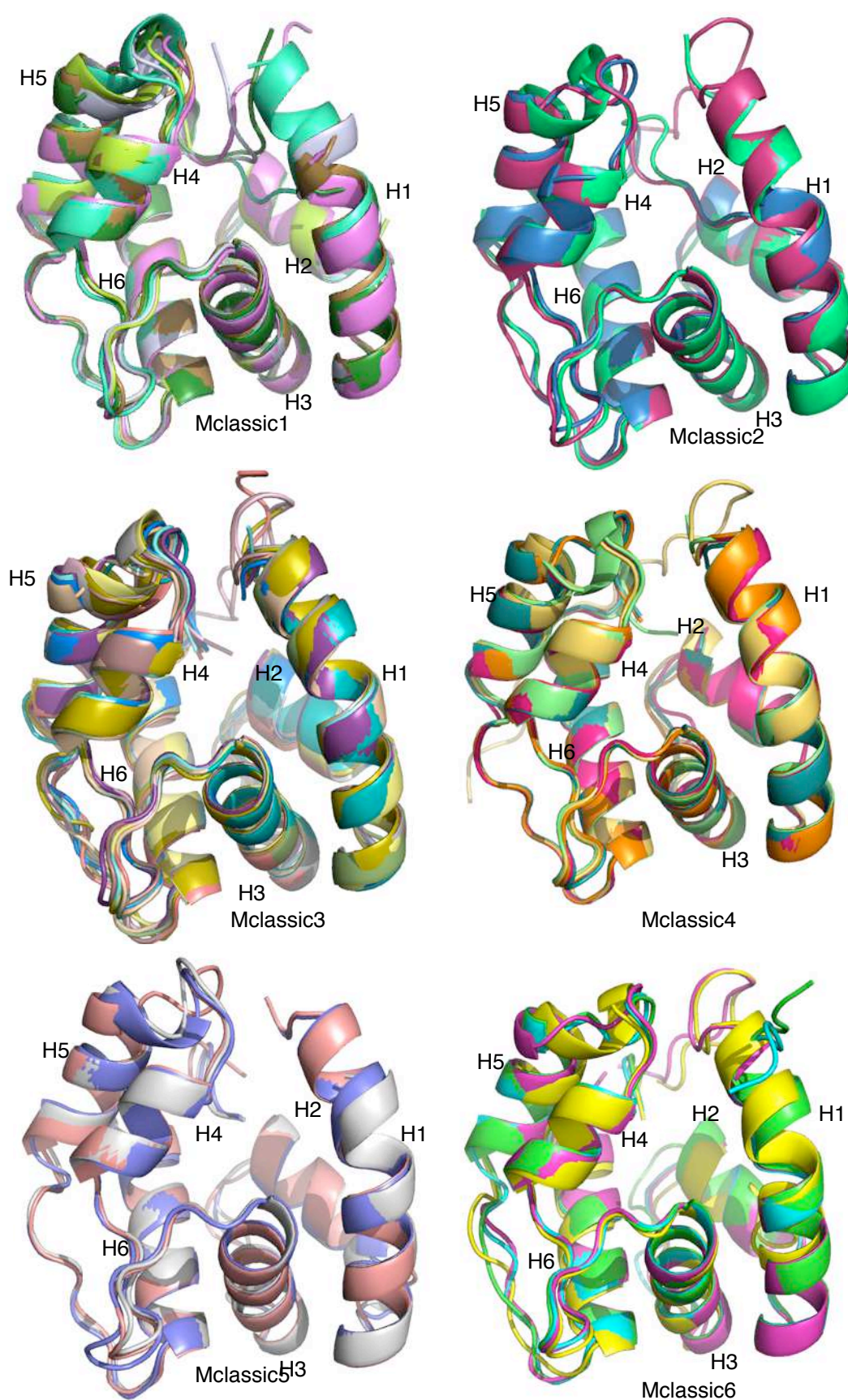


Figure 4.4. Graphical representation of the superposition of models for every cluster of the *Classic* OBPs. Superimpositions were performed using Mustang Software and representations using PyMOL (continued on next page).

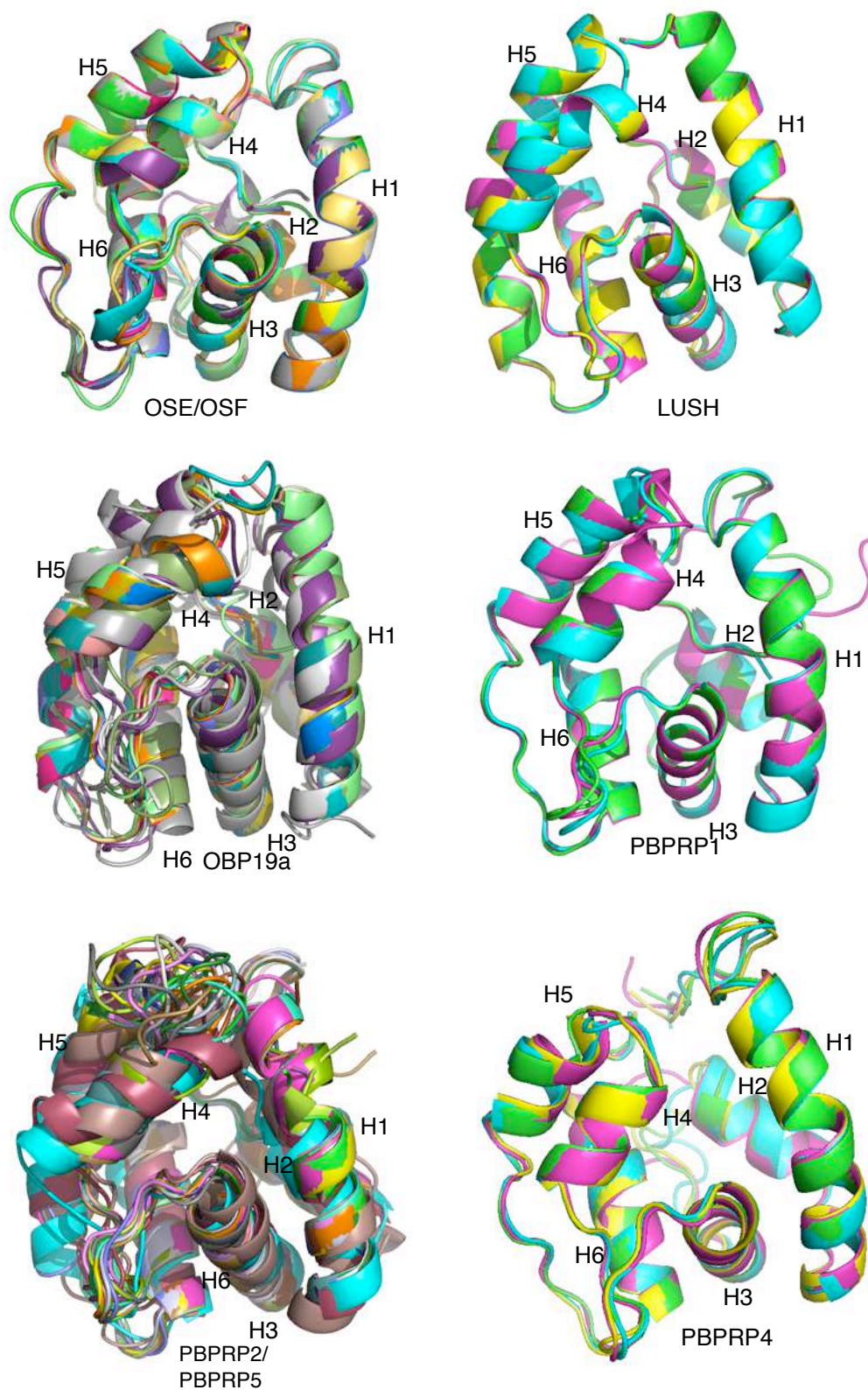


Figure 4.4 (contd). Graphical representation of the superposition of models for every cluster of the *Classic* OBPs. Superimpositions were performed using Mustang Software and representations using PyMOL (continued on next page).

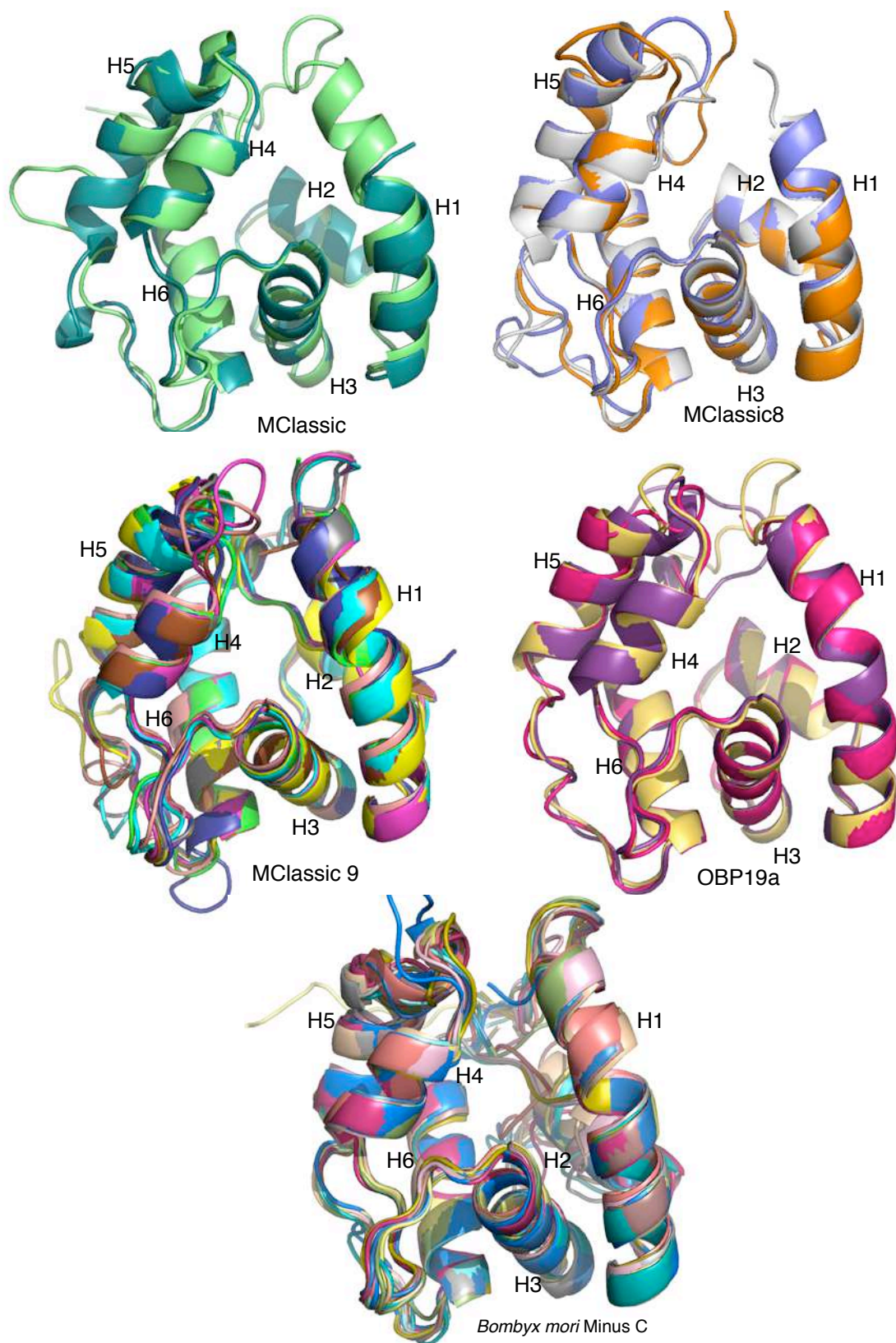
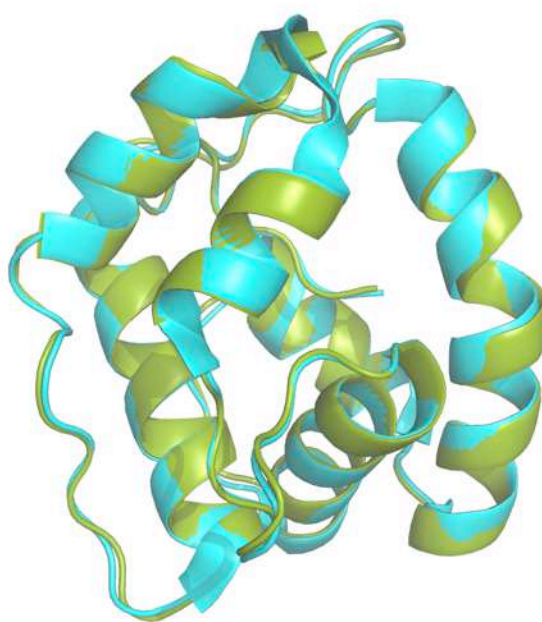
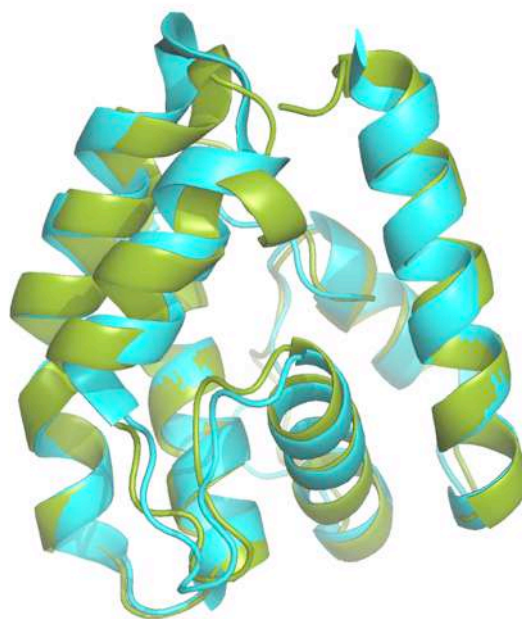


Figure 4.4 (contd). Graphical representation of the superposition of models for every cluster of the *Classic* OBPs. Superimpositions were performed using Mustang Software and representations using PyMOL.



(a)



(b)

Figure 4.5. Comparison between two OBP models (CquiOBP1 and AgamOBP4) and their corresponding crystal structures (PDB:3OGN and PDB:3Q8I) that were later published. (a) Structural superposition of the crystal structure of CquiOBP1 PDB:3OGN (in cyan) with the determined model of CquiOBP1 (in green). The rmsd between the two structures is 0.32 Å. (b) Structural superposition of the crystal structure of AgamOBP4 PDB:3Q8I (in cyan) with the determined model of AgamOBP4 (in green). The rmsd between the two structures is 0.95 Å.

Table 4.1. List of all the *Classic* OBP models built using MODELLER. Show, are their complete sequence length, signal peptide region, model length, template, no of cysteines in the model, sequence identity, Modeler energy, Ramachandran plot - % of fully and additionally allowed regions and the start and end residues of the model.

S.no	OBP	Length			Template	J-score	Cys in model	%identity	Modeler energy	% of fully allowed	% of additionally allowed	Start and end residues
		Full seq	SP	Model								
1	AgamOBP2	157	25-26	125	3k1e	82.3	6	50.8	476.3	96.20	3.80	32-157
2	AgamOBP3	153	29-30	124	3k1e	80.7	6	65.3	457.9	96.30	3.7	30-153
3	AgamOBP4	150	25-26	125	1ooh	83.3	6	40.8	460.5	95.60	4.40	26-150
4	AgamOBP5	156	33-34	125	1ooh	81.3	6	45.5	535.8	94.60	4.50	34-156
5	AgamOBP6	155	33-34	122	2erb	77.0	6	19.7	515.7	91.00	9	34-155
6	AgamOBP7	154	28-29	126	3k1e	79.3	6	29.4	469.8	93.90	6.1	29-154
7	AgamOBP8	176	29-30	123	3k1e	73.3	6	13.3	505.2	89	10.2	44-176
8	AgamOBP9	139	17-18	121	3k1e	73.7	6	21.6	441.2	91.9	5.4	19-139
9	AgamOBP10	131	19-20	126	2erb	61.7	5	19.6	650.1	91.3	7.7	20-131
10	AgamOBP11	192	18-19	121	3k1e	79.7	6	15.0	546.3	90.7	5.6	44-167
11	AgamOBP12	159	27-28	132	1ooh	71.7	6	14.8	614.0	88.3	10	28-159
12	AgamOBP13	149	23-24	126	2erb	76.7	6	19.3	448.3	91.2	7.9	24-149
13	AgamOBP14	188	22-23	125	2erb	79.0	6	12.5	524.4	92.8	6.3	44-159
14	AgamOBP16	147	18-19	123	2wc5	80.3	6	30.8	507.2	91.7	8.3	19-141
15	AgamOBP19	137	no	120	1ooh	77.3	6	30.0	519.6	95.5	4.5	18-137
16	AgamOBP20	142	14-15	128	2erb	75.0	6	20.7	466.2	89.7	10.3	15-142
17	AgamOBP21	131	20-21	112	3k1e	71.3	6	25.2	416.8	91.1	8.9	21-131
18	AgamOBP22	144	21-22	123	3k1e	70.3	5	17.8	576.6	89.6	9.6	22-144
19	AgamOBP23	131	19-20	112	2erb	72.7	6	21.6	401.8	94.3	4.8	20-131
20	AgamOBP24	176	no	125	2erb	61.0	6	18.5	504.3	88.3	11.7	39-163
21	AgamOBP25	142	24-25	118	2erb	75.3	6	17.4	399.3	94.4	5.6	25-142
22	AgamOBP26	131	18-19	113	2erb	74.0	6	18.8	362.5	89.9	9.1	18-131
23	AgamOBP27	134	25-26	109	3k1e	61.3	6	14.0	560.2	88.1	10.9	25-134
24	AgamOBP28	134	16-17	118	2erb	74.7	6	17.5	409.8	89	9.2	17-134
25	AgamOBP29	176	38-39	125	2erb	71.0	6	15.2	528.6	87.7	9.6	39-163
26	AgamOBP62	174	35-36	118	3k1e	56.5	6	11.0	452.0	88.8	9.3	48-166
27	AgamOBP63	135	19-20	116	3k1e	75.3	6	28.3	375.8	92.3	6.7	20-135
28	AgamOBP64	142	29-30	114	3k1e	71.3	6	15.7	433.3	89.2	8.8	30-142
29	AaegOBP1	146	24-25	125	1ooh	72.5	6	41.0	458.9	92.68	4.88	22-146
30	AaegOBP2	141	21-22	119	1ooh	75.3	6	27.7	492.3	90.5	8.6	21-140
31	AaegOBP3	115	no	115	1ooh	70.0	6	29.6	553.6	95.1	4.9	1-115
32	AaegOBP4	145	25-26	120	1ooh	78.0	6	33.3	470.5	95.2	4.8	26-145
33	AaegOBP8	133	16-17	117	2erb	74.3	6	27.7	416.4	94.1	5.9	17-133
34	AaegOBP9	132	20-21	112	2erb	72.7	6	20.7	394.5	94.1	4.9	21-132
35	AaegOBP10	140	25-26	114	3k1e	71.7	6	23.6	453.9	92.4	6.7	26-140
36	AaegOBP11	137	18-19	119	2erb	77.7	6	20.2	1,780.6	92.5	6.6	19-137

S.no	OBP	Length			Template	J-score	Cys in model	%identity	Modeler energy	% of fully allowed	% of additionally allowed	Start and end residues
		Full seq	SP	Model								
37	AaegOBP12	132	18-19	114	3k1e	73.0	6	18.0	376.8	90.7	8.4	19-132
38	AaegOBP13	132	18-19	114	2erb	74.3	6	25.0	351.1	93.1	6.9	19-132
39	AaegOBP14	132	18-19	114	2erb	74.3	6	25.9	363.9	94.1	5.9	19-132
40	AaegOBP15	136	23-24	109	2erb	71.0	6	14.0	410.8	88.7	9.3	24-128
41	AaegOBP17	136	19-20	111	3k1e	70.3	6	21.7	489.9	93.1	5.9	24-132
42	AaegOBP18	136	22-23	114	2erb	71.7	6	20.5	487.2	90.5	9.5	23-136
43	AaegOBP19	145	26-27	119	3k1e	74.7	6	20.7	520.9	89.6	8.5	27-145
44	AaegOBP20	166	24-25	119	2erb	75.7	6	11.4	585.4	88.8	10.3	38-156
45	AaegOBP21	141	18-19	123	1ooh	72.0	6	17.2	509.3	91.2	7.9	18-141
46	AaegOBP22	138	16-17	115	3k1e	74.0	6	19.5	459.2	90.7	8.4	17-135
47	AaegOBP27	149	23-24	126	2erb	80.7	6	28.5	518.1	94	4.3	24-149
48	AaegOBP34	149	24-25	125	1ooh	83.3	6	37.6	434.6	98.3	1.7	25-149
49	AaegOBP35	131	18-19	113	3k1e	73.3	6	19.6	355.5	93.1	6.9	19-131
50	AaegOBP36	152	26-27	126	2erb	80.7	6	32.0	436.5	91.9	8.1	27-152
51	AaegOBP37	148	20-21	128	3k1e	80.3	6	31.4	525.9	92	8	21-148
52	AaegOBP38	140	16-17	124	3k1e	80.7	6	65.6	483.5	95.5	4.5	17-140
53	AaegOBP55	151	24-25	127	1ooh	77.7	6	30.2	545.2	93.9	5.3	25-151
54	AaegOBP57	144	23-24	121	2erb	74.3	6	18.1	460.4	90.3	8	24-144
55	AaegOBP59	166	24-25	118	2erb	75.3	6	11.4	562.0	88.7	9.4	39-156
56	AaegOBP60	142	18-19	124	3k1e	79.7	6	53.3	464.3	94.4	5.6	19-142
57	AaegOBP61	132	18-19	132	3k1e	80.3	6	26.0	409.9	87.69	6.15	1-132
58	AaegOBP65	91	no	91	3k1e	58.3	6	14.8	317.3	91.6	8.4	1-91
59	AaegOBP76	134	no	120	3k1e	76.7	6	18.8	437.8	91.9	7.2	15-134
60	AaegOBP77	138	20-21	119	2erb	73.7	4	19.7	460.9	89.8	8.3	19-138
61	AaegOBP78	117	18-19	117	2erb	74.3	6	17.2	471.7	90.1	9	1-117
62	AaegOBP79	154	20-21	124	3k1e	74.3	6	15.5	1,034.9	86.5	13.5	21-144
63	AaegOBP80	152	20-21	124	3k1e	73.7	6	19.0	1,006.5	79.6	15.9	21-144
64	AaegOBP81	151	23-24	116	3k1e	70.7	6	16.8	546.6	88.3	11.7	24-139
65	CquiOBP1	149	24-25	125	3K1E	82.7	6	87.1	460.0	97.3	2.7	25-149
66	CquiOBP2	146	22-23	124	3K1E	80.7	6	67.8	483.2	96.4	3.6	23-146
67	CquiOBP3	147	18-19	129	2erb	80.7	6	54.1	533.0	91.7	8.3	19-147
68	CquiOBP4	150	18-19	132	3K1E	81.7	6	37.1	516.5	90.8	9.2	19-150
69	CquiOBP5	143	15-16	128	3K1E	81.7	6	37.9	526.1	93.9	5.2	16-143
70	CquiOBP7	136	no sig	136	3K1E	78.0	6	22.6	498.8	93.3	5.8	1-136
71	CquiOBP8	144	23-24	121	1ooh	77.3	6	34.7	480.4	93.3	6.7	24-144
72	CquiOBP9	139	20-21	119	1ooh	79.3	6	25.2	437.8	93.5	6.5	21-139
73	CquiOBP10	132	no sig	125	1OOH	78.0	6	35.0	1,064.3	90.8	7.3	8-132
74	CquiOBP11	144	23-24	121	1OOH	75.3	6	33.1	533.7	91.9	7.2	24-143
75	CquiOBP12	121	22-23	98	1OOH	60.3	5	29.3	453.5	95.5	2.2	24-121

S.no	OBP	Length			Template	J-score	Cys in model	%identity	Modeler energy	% of fully allowed	% of additionally allowed	Start and end residues
		Full seq	SP	Model								
76	CquiOBP13	143	23-24	120	2erb	75.3	6	20.0	448.7	92.5	7.5	24-143
77	CquiOBP14	170	20-21	118	2erb	74.7	6	14.4	492.2	89.6	10.4	43-160
78	CquiOBP15	141	28-29	113	2erb	71.0	6	14.0	420.2	90.1	6.9	29-141
79	CquiOBP16	134	20-21	114	2erb	70.3	6	10.4	395.0	90.5	9.5	21-134
80	CquiOBP17	132	18-19	114	2erb	74.3	6	22.3	347.4	92.3	7.7	19-132
81	CquiOBP18	132	18-19	114	2erb	74.0	6	18.8	391.6	89.5	8.6	19-132
82	CquiOBP19	139	17-18	122	3K1E	79.0	6	16.5	412.1	90.6	8.5	18-131
83	CquiOBP20	131	18-19	113	2erb	73.7	6	22.1	363.6	92.1	6.9	19-131
84	CquiOBP21	139	28-29	111	3K1E	70.7	4	14.8	1,782.6	91.4	8.6	29-139
85	CquiOBP22	131	19-20	112	2erb	73.0	6	20.7	378.6	92.3	7.7	20-131
86	CquiOBP23	136	17-18	119	3K1E	74.3	6	18.3	455.0	88.5	10.6	18-136
87	CquiOBP24	137	23-24	114	3K1E	72.3	6	23.6	444.8	92.4	6.7	24-137
88	CquiOBP25	121	16-17	105	3K1E	66.0	6	16.0	427.1	84.7	13.3	17-121
89	CquiOBP26	119	15-16	104	3K1E	64.7	6	20.2	395.8	90.5	8.4	16-119
90	CquiOBP27	126	21-22	105	2erb	63.7	6	14.3	353.8	90.5	9.5	22-126
91	CquiOBP28	150	20-21	130	2erb	76.7	6	18.5	505.8	89	7.6	21-150
92	CquiOBP29	130	no sig	130	2erb	79.3	6	13.8	529.1	88.9	9.4	1-130
93	CquiOBP30	143	20-21	123	3k1e	75.0	6	18.1	430.3	89.2	9.9	21-143
94	CquiOBP31	124	16-17	108	2erb	68.3	6	16.0	390.3	90.9	9.1	17-124
95	CquiOBP32	126	18-19	108	1ow4	68.0	6	20.6	425.6	93.8	3.1	19-126
96	CquiOBP33	124	19-20	105	2erb	65.0	6	20.0	351.7	89.7	10.3	20-124
97	CquiOBP34	116	no sig	116	2erb	72.7	6	15.0	404.2	90.5	8.6	1-116
98	CquiOBP35	126	18-19	108	1ow4	65.7	6	21.3	445.8	92.8	5.2	19-126
99	CquiOBP36	146	18-19	128	2wc5	77.0	6	17.5	508.7	93	7	19-146
100	CquiOBP37	135	no sig	128	3K1E	78.7	6	15.4	440.0	92	8	1-128
101	CquiOBP38	137	20-21	117	2erb	70.0	6	16.5	386.7	94.2	4.8	21-137
102	CquiOBP39	126	18-19	108	2erb	66.0	6	17.6	380.4	87.6	12.4	19-126
103	CquiOBP40	107	no sig	107	2erb	67.3	6	14.3	354.2	90.7	9.3	1-107
104	CquiOBP41	98	no sig	98	2erb	60.3	5	20.2	354.5	90.7	8.1	1-98
105	CquiOBP42	111	no sig	111	2erb	66.0	6	19.2	367.8	91.2	7.8	1-111
106	CquiOBP43	155	no sig	123	3K1E	78.3	6	18.3	441.5	92.2	6	9-131
107	CquiOBP44	147	20-21	118	3K1E	70.7	6	22.4	532.3	91.3	8.7	21-138
108	CquiOBP46	150	22-23	128	2erb	75.3	6	12.1	525.5	90.7	9.3	23-150
109	CquiOBP51	144	no sig	124	2erb	77.3	6	13.6	493.6	91.7	8.3	10-133
110	CquiOBP52	143	22-23	124	3K1E	70.0	6	15.0	507.4	93.5	6.5	20-143
111	CquiOBP53	145	19-20	130	1gm0	74.0	5	19.0	540.0	92.4	6.8	1-130
112	CquiOBP54	143	19-20	122	2erb	73.7	6	20.0	423.3	91	7.2	49-170
113	CquiOBP56	214	no sig	111	2wc5	67.7	5	18.9	451.1	91.1	7.9	101-214
114	CquiOBP57	126	no sig	126	1OOH	76.7	6	22.6	520.9	94.7	4.4	1-126

S.no	OBP	Length			Template	J-score	Cys in model	%identity	Modeler energy	% of fully allowed	% of additionally allowed	Start and end residues
		Full seq	SP	Model								
115	CquiOBP58	113	no sig	113	3bjh	82.0	6	20.9	454.3	91.2	7.8	1-113
116	CquiOBP59	138	18-19	128	3K1E	82.5	4	13.9	560.5	88.4	9.9	11-138
117	CquiOBP60	138	17-18	129	2erb	79.3	4	9.9	511.5	90.4	9.6	10-138
118	CquiOBP61	120	no sig	120	2erb	69.7	4	17.0	457.2	89.7	10.3	1-120
119	CquiOBP62	181	no sig	126	3dxl	84.5	4	17.7	1,166.6	85.2	11.3	56-181
120	CquiOBP63	206	no sig	132	3ogn	80.3	6	14.0	528.9	86.92	6.92	32-163
121	CquiOBP64	136	20-21	124	2erb	80.3	4	16.2	446.0	93.8	6.2	13-136
122	CquiOBP65	136	17-18	126	2erb	83.0	4	15.8	466.5	89.5	9.6	11-136
123	CquiOBP66	130	no sig	123	3K1E	81.5	4	16.1	421.0	92.9	6.2	8-130
124	CquiOBP67	119	118.34	119	2erb	76.8	4	13.0	449.7	89	10.1	1-119
125	CquiOBP68	137	no sig	125	3B87	80.5	4	15.3	565.9	91.2	7	13-137
126	CquiOBP69	122	no sig	122	2erb	77.8	4	16.4	501.5	92.8	5.4	1-122
127	CquiOBP69	136	19-20	125	2erb	80.3	4	20.3	498.6	90.5	8.6	12-134
128	Cquiobp70	134	17-18	134	3K1E	76.3	4	18.6	431.2	93.5	6.5	1-134
129	CquiOBP72	98	no sig	98	2qev	68.5	3	12.2	373.1	9	7.8	1-98
130	CquiOBP73	132	17-18	123	2erb	82.8	4	9.1	1,042.9	87	11.3	10-132
131	CquiOBP74	128	20-21	116	3K1E	69.7	4	11.2	510.9	89.8	8.3	13-128

5

Towards unravelling the molecular mechanism underlying the functioning of an OBP through molecular dynamics simulations

5.1. Introduction

With the current knowledge available on Odorant Binding Proteins (OBPs) of insects, it is now strongly believed that OBP serve as primary transporters involved in importing the odorant molecule from the sensillum lymph to the neuronal membrane where they are presented to the ORs for receptor activation. For an OBP to function as a carrier and for it to play additional putative roles in odor discrimination, receptor activation and odorant deactivation, its uptake/release mechanisms need to be individually tuned (Steinbrecht, 1998). How this is to be achieved is yet to be elucidated. As for OBPs in general, a crucial and yet unsolved question is the mechanism of ligand release.

The first hint on this came from the fact that the tertiary structure of BmorPBP, a transporter for the pheromone bombykol through the sensillar lymph of the antennae to the pheromone receptor in *Bombyx mori*. The pheromone binding protein is sensitive to pH changes and it undergoes dramatic conformational transition between pH 5.0 to 6.0 described by the analysis from circular dichroism and fluorescence spectroscopy (Wojtasek and Leal 1999). This pH-dependent conformational change was later predicted to occur at the loop from residues 60-69, a His-rich loop between helices 3 and 4, after the crystal structure of the same protein was deciphered at a pH 8.2 (Sandler et al. 2000). It was related to the protonation of three His residues seen in this loop at low pH. In the same year, NMR spectroscopic studies aimed at investigating the changes as a function of pH in solutions of BmorPBP (Damberger et al. 2000, showed that it undergoes a conformational transition between pH 4.9 and 6.0. The protein was believed to exist in an “acid/A form” at a pH below 4.9 and a “basic/B form” above pH 6.0 (Damberger et al. 2000). The NMR structure

assignments to the acid/A form of the protein showed the protein was found to have a tightly packed arrangement of seven helices, in contrast to the crystal structure of the same protein which had only 6 helices at pH 8.2 in the structure solved earlier (Sandler et al. 2000). The difference was observed in the C-terminal dodecapeptide, which in the case of the pheromone complex at pH 8.2, is an extended conformation located on the protein surface. On the contrary, this region forms a regular helix located in the pheromone binding site in the case of the unliganded form of BmorPBP^A (Horst et al. 2001). Otherwise, the NMR structure of the BmorPBP^B was found to be more closely related to the crystal structure - with a disordered C-terminal end outside the binding pocket of the protein (Lee et al. 2002). Subsequently, in 2005, the crystal structure of the apo (bound) form of BmorPBP^B was determined at a pH 7.5. The crystal structure of the unliganded BmorPBP^B most closely resembles the NMR structure of BmorPBP^A, where the C-terminal tail forms an ordered helix occupying the binding pocket. It was hence hypothesized that the BmorPBP can exist in two different conformations at neutral pH. Thus the conformational changes observed are not only pH-sensitive but also sensitive to the presence or absence of ligand. An hypothetical model for pheromone release was proposed based on these observations (Figure 5.1) (Lautenschlager et al. 2005). Molecular dynamics studies to analyze the parts of the same protein involved in such mechanisms were also carried out by certain groups. Nemoto et al. 2002 showed, from their 1 ns simulation, that the loop 60-69 was the most flexible region of the protein and its role as a flexible lid for the binding pocket in the pheromone binding phenomenon. Subsequent molecular dynamics simulations for an extended period of 50 ns showed that in addition to this loop reported previously, the N-terminal (1- 24) and C-terminal (125-137) and the loop (99-106) also showed remarkable flexibility (Grater et al. 2006). In the same study, it was also observed that the bombykol undergoes a partial unbinding in the binding pocket. This was further analyzed subsequently by replica exchange, essential dynamics and force probe molecular dynamics (Grater et al. 2006). The results suggested two opposite dissociation routes for bombykol, one of which runs along a flexible front lid and the other along the termini at the back. These two routes were stated to be physiologically relevant from calculated forces and energies.

It was followed by studies on pheromone binding protein 1 from the wild silkmoth *Antheraea polyphemus* (ApolPBP1), where this protein was also shown to undergo a pH-dependent structural transition. HSQC (heteronuclear single quantum coherence) spectra recorded at pH 4, 5, 6 and 7 showed that their patterns varied significantly between pH 5 & 6, whereas patterns between pH 4 & 5 and pH 6 & 7 were more similar, showing that there could be conformational transition between pH 5 and 6 (Mohanty et al. 2004). The NMR structure of ApolPBP1^A shows that the

protein shares the same global fold as BmorPBP^A consisting of 7 helices with the helix7 occupying the binding pocket. It was also noted that the loop between helix 3 and 4, which shows conformational heterogeneity, as compared between BmorPBP^A and BmorPBP^B was also found in ApolPBP1^A structures. It was hence suggested that this protein could also indulge in a similar mechanism of ligand binding and release involving the loop between helix3 and helix4 (Damberger et al. 2007).

While these two pheromone binding proteins were found to undergo a pH-dependent conformational change, the odorant binding protein from *Drosophila* LUSH was found to undergo a ligand-dependent conformational change. Though the same was suggested also for the *Bombyx mori* PBP, the evidence for this was not as direct as observed in the case of odorant binding protein LUSH. LUSH bound to cVA forms an interaction with Phe121 at the C-terminal end of the protein which is found inside the binding pocket of LUSH. This is unlike the BmorPBP and LUSH structures, where they are located outside the binding pocket in ligand bound forms of the protein. It is also interesting to note that the C-terminal segment of LUSH is shorter than the C-terminal segment of those observed in the BmorPBP and ApolPBP limiting their possibility to form a helical structure. The Phe121 interaction with cVa appears to mediate specific conformational shifts to residues Gln120 and Asp118 which results in the disruption of a salt bridge between Asp118 and Lys87 which is otherwise present in the apo and alcohol bound forms of LUSH (Kruse et al. 2003; Thode et al. 2008). Other conformational changes were observed at the loop connecting helices 2 and 3, which finally results in an invagination on the surface of the protein of ~150 Å² that is open to solvent that could potentially function as a recognition site for binding partners. It was thus suggested that the LUSH-cVA complex directly activates the ORs and such pheromone-induced conformational change would be detectable if the neuronal receptor complex is specifically tuned to that conformation increasing the sensitivity and specificity of the pheromone detection process.

No experimental or computational data relating to such analysis, oriented towards the ligand binding and release mechanism of odorant binding proteins in mosquitoes, have been pursued so far kindling the curiosity of the mechanism adopted by them. However, hypothesis on the possible mechanisms have been proposed for OBPs from *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* with reference to the structures discussed above. The C-terminal end of the mosquito odorant binding proteins is short as in the case of LUSH and is described as the wall of the binding pocket by the authors. In all the three structures, it was seen that C-terminal end was held inside the binding pocket by hydrogen bonds between the terminal residue of the C-terminal and residues in the helix 1 and helix 3. In addition, other hydrogen bonds were found to play an

important role in holding this C-terminal “lid” in place namely Arg5-Arg7, Arg6-Arg42, Asp118-His121, Asp118-Lys120 and Asp7-Tyr10, as in the case of *Anopheles gambiae* and *Aedes aegypti*. It is likely that the some of these interactions could be disrupted at low pH, leading to the C-terminal residues to move away from the binding pocket thus opening the “lid” and paving way for the ligand. Circular dichroism studies on the effect of pH in the case of *Aedes aegypti* showed significant change in the near-UV CD spectra, indicating a change in the tertiary structure while the far-UV CD spectra remained indistinguishable suggesting an intact secondary structure. In the case of odorant binding protein from *Culex*, the hydrogen bond between the C-terminal, Helix1 and Helix 3 was proposed as a “pH sensing triad” which could be disrupted at lower pH conditions and therefore displacing the C-terminal end from the binding pocket (Figure 5.2). In studies of CquiOBP1 using NMR, it was also observed that the residues in helix5 exhibited exchange-broadened NMR resonances suggesting that this region may undergo a conformational exchange. It was proposed that these conformational fluctuations in helix 5 may function as a gate to help create an opening to allow the entrance of its ligand MOP inside the protein (Mao et al. 2010).

Based on these assumptions of a pH dependent conformational change in the case of mosquito OBPs, we formulated simulation experiments at varying pH to analyze the effect of pH on the conformational flexibility of the protein in solution. The recently available structure of CquiOBP1 (PDB ID:3ogn; in complex with an oviposition pheromone MOP/3OG (C18 H32 O4) from *Culex* was used as a starting structure for this analysis. Four different experimental conditions were postulated : at two different pH conditions 8.2 (native/high pH) and 4.0 (low pH) and in the presence and absence of the ligand each. Each experimental setup was duplicated using ransom seed numbers making a total of 22 individual 50 ns simulations. These experiments were designed to further analyze the hypothesis proposed on the ligand binding and release mechanism of mosquito OBPs described previously in this chapter and provide new insights on the same.

5.2. Methodology

GROMACS 4.5.3 (Van Der Spoel et al. 2005) molecular dynamics package was used with the OPLS-AA forcefield (William L. Jorgensen 1996) for the simulations. Simulations were carried out on the *Culex* OBP (CquiOBP1) crystal structure PDB ID:3OGN with and without ligand at two different pH conditions pH8.2 (native pH) and pH4.0 (low pH). The pH change was mimicked by the protonation of histidine residues based on the pka values predicted by PROPKA (Li et al. 2005). 8 of the 9 histidine residues in the structure were protonated at low pH. His111 was not protonated

since it was found more buried in the structure and it had a low pKa value of 2.93. OPLS-AA force field parameters for MOP/3OG (C18 H32 O4) (1S)-1-[(2R)-6-oxotetrahydro-2H-pyran-2-yl]undecyl acetate were built by careful manual chemical intuitions. Bonded and non-bonded parameters for the ester and aromatic ester ring like groups of the ligand was obtained from Price et al. (2001) optimized for ester groups. The parameters for the aliphatic side chain and methyl groups were adapted from the OPLS-AA force field. The bond stretching, angle bending and dihedrals and impropers were also adapted from the same sources respectively. A zero total charge was applied on the ligand. The starting structures of every simulation were solvated in cubic periodic water boxes with a 1.4 nm solute-wall distance. The system was energy minimized twice before and after the addition of ions. The native pH simulation system was neutralized using 7 Na⁺ atoms and the low pH system was neutralized using one Cl⁻ ion. The size of the system was approximately 28,500 atoms. The system was equilibrated for 100ps with position restraints prior to the 50 ns production run. A two femtosecond time step was used for integration of the equations of motion. Solute and solvent with the ions were coupled to a reference temperature bath at 300K with a coupling constant T of 0.1ps. The pressure was maintained by weakly coupling the system to an external pressure bath at one atmosphere with a coupling constant P of 1.0ps. Non-bonded interactions were calculated using twin range cutoffs of 0.8 and 1.4 nm. Long-range electrostatic interactions beyond the cutoff were treated by PME Electrostatics with an order of 4 and fourier spacing of 0.12. The LINCS algorithm was used for bond length constraints in conjunction with dummy atoms for the aromatic rings and amino group in side chains. The same experimental setup was duplicated using random seed numbers each 50ns in length for each of the different starting structures in order to avoid any bias on the interpretation of the results. 22 simulation systems were setup as detailed in Table 5.1.

5.3. Results

22 explicit-solvent MD simulations of CquiOBP1 of 50ns each were computed. Each simulation began with a protein conformation based on the 1.3Å resolution crystal structure (Mao et al. 2010). Four different simulations systems were setup; (i) high pH (8.2) without ligand; (ii) low pH (4.0) without ligand; (iii) high pH (8.2) with ligand; (iv) low pH (4.0) with ligand. The conformational stability of the various systems were assessed by the drift of the protein from the initial structure using RMSD of C^α atoms (Figure 5.3). Throughout each simulation, the overall conformation of the protein remained close to the crystal structure. The systems at low pH (pH 4.0) showed higher variations in terms of side chain flexibility (Figure 5.4b, d) compared to the systems

at native pH (pH 8.2) which was measured by the RMS fluctuations of individual residues (Figure 5.4a, c). The N-terminal residues (1-10) of the protein were found to have the highest fluctuation and hence contributed to high RMSD ranges which could be a result of their location outside the protein as an extended part giving them the maximal flexibility space in a water box. This region was excluded from other analysis on the systems. Since the starting structures with ligand depict a more realistic and acceptable starting structure, further analysis described in this chapter is restricted to the simulations with ligand at different environmental conditions on a note of precaution. However at this level the results with and without the ligand are comparable, were similar RMS fluctuations are observed for the simulations systems with and without ligand (Figure 5.4).

Apart from the N-terminal segment of the protein it was also observed that the residues from 63-71 which is the loop connecting helix3 and helix4 showed an increased flexibility compared to other regions of the protein in all the four simulation systems. Interestingly, in addition to this, fluctuations were observed from residues 77 to 124 specific only in low pH systems which involve helix4, helix5, and helix6. This proved to be a critical observation of this analysis which is further explained in detail. The cysteine residues in this region, involved in the disulphide formation, do not fluctuate like the other residues in this region and hence there is no loss of secondary structure observed. A very intriguing observation is the amount of residue fluctuations in this region did not show comparable fluctuations in the high pH condition suggesting a pH-dependent conformational change does exist in the case of CquiOBP1.

5.3.1. pH sensing triad

We confirm that the C-terminal segment of Cquiobp1 is located inside its binding pocket and is stabilized by hydrogen bonds between residues the terminal residue, and the residues in helix1 and helix3. As hypothesized in the previous work (Mao et al. 2010, the disruption of this hydrogen bond network at low pH would destabilize the C-terminal loop and displace the C-terminal from the central cavity facilitating the release of the ligand. Similar hydrogen bond triad in the C-terminus of the protein is also described in the the structures of *AgamOBP1* and *AaegOBP1*. This “pH sensing lock” (Mao et al. 2010) that clamps the C-terminus was observed closely throughout the simulation .The distances of the carboxylate group of Val125 with the hydroxyl group of Tyr54 and δ -nitrogen of His 23 were monitored throughout the low pH simulation systems (Figure 5.5). It was observed that the distance between His23 and Val125 did not vary from the

typical hydrogen bond distance range in any of the simulations, suggesting a strong interaction between these two residues. However, the distance between Tyr54 and Val125 was found to be more fluctuating, suggesting it to be a rather weak bonding.

5.3.2. Loop between helix 3 and 4

It was further observed that the loop 63 to 74 between helix3 and helix4, which was shown to undergo a pH dependent conformational change in the case of *Bombyx mori* (BmorPBP1) and *Apis mellifera* (ApolPBP1), showed considerable flexibility here in the case of CquiOBP1 as well than the other residues. It was interesting to note the presence of a number of charged residues and hydrogen bonds in this loop which could critically be attributed to several pH-dependent conformational changes (Table 5.2). This loop interacts with helix6 *via* a salt bridge between Asp70 & Lys106 and a hydrogen bond between Lys75 and Val65 with helix4. The distances between the NZ and CG atoms of Lys106 and Asp70 were monitored throughout the simulation. It was observed that the distance increased in the case of low pH simulations, but remained unaffected in the case of high pH simulations (Figure 5.6a). This indicates a loss of interaction between Asp70 and Lys106 in the case of low pH. Similarly, the interaction of the loop with helix4 was monitored by measuring the distance between NZ and C atoms of Lys75 and Val65 (Figure 5.8a). The distance between these residues increased in the case of low pH systems, while the distance in the case of high pH remained within hydrogen bonding distance indicating a loss of this interaction between Lys75 and Val65 in a low pH environment. It was, however, interesting to notice a new salt bridge formed between Asp67 and His60 on helix3 in the case of low pH simulations. The distance between the CG and ND1 atoms of these residues was monitored in the case of these residues throughout the high and low pH simulations (Figure 5.6b). The distance between these atoms reduced in the case of low pH compensating the loss of interaction between Asp70 and Lys106. Thus, in the case of low pH, the loop loses its interaction with helix4 and helix6 and forms new interactions with the helix3 which induces a change in the conformational state of the loop. The three possible bonds of Asp 66 OD1 with Asn68, Gly69 and Asp 70 N and the hydrogen bond distance between NZ atom of Lys63 and the main chain O atom of His60 remains unchanged in both the simulations. This helps in maintaining the overall fold of the loop.

5.3.3. Change in interaction patterns of helix4 and helix5

The change in the conformation of the loop causes a change in the interaction pattern of certain residues in helix4 and helix 5. Interestingly this part of the protein also indicated a high RMS fluctuations in the case of the low pH simulations. A detailed analysis on this region of the protein in the case of low pH simulations indicated changes in interaction pattern of certain residues. During the course of simulations, Asp70 of loop3, which was previously interacting with helix6, forms a new interaction with His72 (measured as a distance between ϵ -nitrogen of His72 and the carboxylate carbon of Asp70 (Figure 5.9a). A correlated increase in the distance between His72 and Glu74 was noticed and is represented as a measure of distance between the HE1 and OE & atoms of His72 and Glu74, respectively (Figure 9b). A new interaction of the δ -nitrogen of His 77 with the carboxylate group of Glu74 (Figure 9c) was formed. As a consequence, the interaction between the main-chain oxygen of His85 and ϵ -nitrogen of His 77 is disrupted. These sequential changes in the interaction patterns in helix4 and helix5 changes the orientation of the side chain of His77, which in turn, appears to cause a change in the surface area of the entrance of the ligand due to the presence of a bulky aromatic group at the opening (Figure 5.10).

5.3.4. Binding pocket and movement of MOP

Partial unbinding of MOP from the beginning of the tunnel embarked by helix4 and 5 was noticed which could be a result of a series of change in the hydrogen bonding network of the helix4 and helix5 described earlier. The entrance of the hydrophobic tunnel found between helix 4 and 5 tends to close upon the partial unbinding of the ligand. Interestingly the partially unbound ligand moves towards the opening located at the convergence of helix5, helix1, helix3 and the loop3 that undergoes a conformational change. Upon tracking the various interactions of MOP at the end of simulations, an extension of the hydrophobic tunnel was observed, contributed by residues Ala18, His23, Leu58, Phe59, Ala62, Val64, Lys75, Met 84, Met89, Leu96, Val125 and Leu124 apart from the ligand binding residues at the initial state of the ligand in the crystal structure (Figure 5.11). This clear representation of a hydrophobic tunnel, towards the opening at the convergence of helix 3 helix4, helix1 and loop 3, suggests that this could be the exit of the ligand. The same tendency to close the entrance of the binding pocket is not seen in the case of the high pH simulations (Figure 5.11).

5.3.5. Essential dynamic analysis

To analyze the effect of the change in the interaction patterns on the protein motions, we extended our analysis using the essential dynamics analysis or principal component analysis. The trajectories from 0-50 ns of all the four low-pH simulations with the ligand were used to construct a covariance matrix. High positive peaks at the regions described, involving the loop, helix 4 and C-terminal end of helix6 indicating a correlated movement of the loop, helix3 and the C-terminal part of the helix6 supports the above description (Figure 5.12c). Projection of the combined trajectories on the first five eigenvectors which could represent the main motions of the protein indicated that, simulation at a low pH, showed higher motions than the simulation at high pH conditions (Figure 5.12a,b). The observations made based on the essential dynamic analysis can be correlated to the change in the interaction patterns described above. The loss of a salt bridge between the terminal residue of the loop and helix6 and the loss of a hydrogen bond between Val65 and Lys75 causes a conformational drift of the loop causing a anti-clockwise flip of the loop, with respect to helix3 and helix4 (Figure 5.7c). This loss of interaction makes the side chains of Asp67 and Asp70 available for new salt bridges with His60 in helix3 and His72 in helix4, respectively, where the interaction between Asp67 and His60 tilts the loop toward helix3. Thus, the conformational change of loop is mediated by a correlated change in the interaction pattern, mainly involving the charged residues in the loop. With the loss of the hydrogen bond between Lys75 and Val65, Lys75 is made available for interaction with the ligand. This also results in a noticeable change in the orientation of helix4 increasing the distance between helix1 and helix4. The interaction of His 72 of helix 4 with Asp67 increases the distance between His72 and Glu74 which makes the side chain of Glu74 of helix 4 for a salt bridge with His77 which causes the closure of the entrance of the binding pocket.

5.4. Discussion

5.4.1. CquiOBP1 undergoes a pH dependent conformational change

Globular proteins in general are capable of reacting to changes in environmental conditions such as temperature, pH, ligand concentration and changing their conformation that is facilitated by a number of interdependent cooperative interactions embedded in the rather complicated steric arrangement of their polypeptide chain. OBPs have been previously described to undergo conformational changes mediated by both pH change and ligand binding for their primary roles in olfaction. The pH change at the vicinity of the dendritic membrane was described to induce necessary conformational changes in the protein which releases the ligand to activate receptors,

which then directly activates the corresponding odorant receptors (Wojtasek and Leal 1999; Damberger et al. 2000; Sandler et al. 2000; Horst et al. 2001; Zubkov et al. 2005). Alternatively, it has also been described that a few OBPs are capable of activating the receptor in complex with the ligand without the release of the ligand (Campanacci et al. 2001; Bette et al. 2002; Mohl et al. 2002; Honson et al. 2003; Kruse et al. 2003; Xu et al. 2005). This is facilitated by conformational changes of a protein caused by ligand binding which are further recognized by their receptors (Laughlin et al. 2008). Evidence for a pH dependent conformational change has been described in AgamOBP1, and AaegOBP1 using CD spectra associated with loss of affinity for ligand in the case of AgamOBP1 (Wogulis et al. 2006) and change in the near-UV CD spectra indicating a change in the tertiary structure in the case of AaegOBP1 (Leite et al. 2009). It has also been very recently described that the ligand-induced conformational ordering can play a key role in regulating the heteromeric interactions between OBPs using the structure of AgamOBP4 (Davrazou et al. 2011). CquiOBP1, which is a ortholog of AgamOBP1 and AaegOBP1, was also predicted to undergo a pH-dependent conformational change. With the assumption that CquiOBP1 undergoes a pH- dependent conformational change, we simulated the protein with the ligand at different pH conditions. The change in pH was mimicked by changing the protonation states of histidine residues of CquiOBP1. The change in pH was found to have minimal effect on the overall stability of the protein with intact secondary structure, corresponding to the results observed for AaegOBP1 using CD spectra. However, the results indicated higher RMS fluctuations observed in certain regions of the protein specific to low pH simulations, which is completely absent in the high pH simulation systems. This strongly suggests CquiOBP1, similar to AgamOBP1 and AaegOBP1, is prone to undergo conformational changes in response to a change in pH without loss of structure.

5.4.2. Does the previously hypothesized “pH sensing triad” of the C-terminal carboxylate contribute to conformational changes seen in the case of low pH simulations?

The C-terminal end of the BmorOBP1 (odorant binding protein from *Bombyx mori*) is described to play an important role, undergoing a significant conformational change at low pH, where the C-terminal otherwise found outside the binding pocket folds itself into a helical structure inside the binding pocket of the ligand. But, in the case of few OBPs like LUSH, Amel-ASP1, AgamOBP1, AaegOBP1 and CquiOBP1, the C-terminal end is too short to form a helix that will occupy the binding pocket. However, in the case of AgamOBP1, AaegOBP1 and CquiOBP1, it was described to form a wall of the binding pocket held in place by a hydrogen bond triad formed between the carboxylate group of Val125 with the hydroxyl group of Tyr54 and δ -nitrogen of His23

and few other interactions. It was hypothesized that this hydrogen bond triad could be a pH-sensing triad, which, upon contact with a low pH environment could disrupt, releasing the C-terminal end from the central cavity. A close examination of the distances of the atoms involved in the triad strikingly showed that the hydrogen bond between Val125 and His 23 remains unaffected throughout the low pH simulations. Thus, the conformation of the C terminal end remains buried in the central binding pocket close to the crystal structure held in place by this hydrogen bond. Thus, we propose that the C-terminal end of the protein is not directly responsive to a pH change. In support of this, it was found that the crystal structure of AgamOBP4 crystallized at pH 6.8 indeed showed the C-terminus to be a part of the binding pocket. This opens up a possibility of having other possible regions of the protein involved in the ligand binding and release mechanisms.

5.4.3. Loop3 of CquiOBP1 undergoes a major conformational change

In addition to the C-terminal end, a histidine-rich loop between helix3 and helix4 has been implicated in ligand binding as a flexible lid. This loop is observed to adopt different conformational states in the structures of OBPs from *Bombyx mori* and *Antheraea polyphemus* pheromone binding proteins (Sandler et al. 2000; Grater et al. 2006; Damberger et al. 2007). It was noticed that this loop in CquiOBP1 bears a number of charged residues and adopts a conformational shift in the case of low pH simulations. At high pH conditions, this loop interacts with helix6 and helix4 but in the case of low pH conditions adopts a new conformation. It undergoes an anti-clockwise rotation and interacts with helix 3, losing its previous interactions with helix6 and helix4 (Figure 5.7). This change in the interactions between the helical segments of the protein and the loop is facilitated by change in the interaction patterns of two salt bridges and two hydrogen bonds. The loss of a salt bridge between helix6 and the loop (Asp70-Lys106) in the case of the low pH simulations is compensated by a new salt bridge between the loop and helix3 (Asp67-His60). The hydrogen bond between the loop and helix4 (Lys75-Val65) is replaced by another hydrogen bond between the same parts of the protein, but involving alternate residues (Asp70 and His72). Hence, we hypothesize that in CquiOBP1, this loop between helix4 and helix5 is directly affected by a change in the pH rather than the C-terminal end of the protein. This change in the conformational state of the loop can play important functional roles possibly providing new insights into the ligand binding and release mechanisms of CquiOBP1. It is also interesting to note that this particular loop shows an overwhelming conservation of charged residues among all the classic OBPs in the

mosquito genome (Chapter 2: supplementary material), further encouraging the fact that the role of this loop can be extended to the other classic OBP members.

5.4.4. Concerted change in interaction patterns following the conformational change of the loop

It was interesting to note that the new conformational state adopted by the loop, observed during our simulations, was followed by coordinated sequential changes in the interaction pattern of certain residues in helix4 and helix5, which in turn, causes a change in the surface of the protein. This change in interaction patterns accounts for the high RMS fluctuations observed initially in the analysis specific to the low pH simulations. The residues in helix4 and 5 in CquiOBP1 form a part of the hydrophobic tunnel involving the binding of MOP. NMR study on the CquiOBP1-MOP complex at pH 7.0 described in Mao et al (2010) indicated that a long stretch of amino acid residues in helix $\alpha 5$ exhibited exchange-broadened NMR resonances suggesting that this region may undergo some type of conformational exchange. It was further proposed that this conformational fluctuation in helix5 may function as a gate to create an opening to allow the entrance of the binding pocket. The current results support the previously assumed hypothesis involving conformational fluctuations of the helix5 and the observed change in the surface of the protein. The current analysis suggests these conformational fluctuations are preceded by a change in the conformational state of the loop3 between helix3 and helix4.

5.4.5. Hypothesized exit of the ligand

Ruling out the option of the unbinding of the C-terminal end from the binding pocket for the release of the ligand, the partial unbinding and movement of the MOP towards the opening at the convergence of helix1, helix3, helix4 and loop 3 stimulates the idea of this being a possible exit route of the ligand. This can be a coordinated effect, corresponding to a change in the conformational state, which causes the movement of helix4 increasing the distance between helix1 and helix4. However, a complete unbinding of MOP was not observed in the simulation. If unbinding should occur, it would require longer simulations. This opening has also been described in AgamOBP4 crystallized at a rather low pH 6.5 and it is described to be the binding site for AgamOBP1. This leaves us with a speculation whether MOP will be released at this exit site or if ligand-induced conformational changes could occur at this end of the protein inducing the binding of other OBPs. Further extended simulations may provide answers to this.

5.5. Conclusion

Odorant binding proteins have proved to be evident to conformational changes mediated by either change in pH or ligand binding. Equal amount of data support both the perspectives. The current analysis on CquiOBP1 and its correspondence with previous experimental analysis strongly suggest that it is more likely to undergo a pH-dependent conformational change. The current prevailing hypothesis involves the release of C-terminal loop from the binding pocket facilitating the release of the ligand. However, in contrast to the previously proposed hypothesis, we propose that the C-terminal loop is not directly affected by a change in pH. An alternate role of a loop between helix 3 and helix4 in this role is described in this study. It is suggested that the loop3 undergoes a change in its conformational state which directly affects some of the interaction patterns between helix4 and helix5. Conformational fluctuations of helix5 have also been previously observed for this protein using NMR analysis supporting the newly provided hypothesis on the ligand binding and release mechanism.

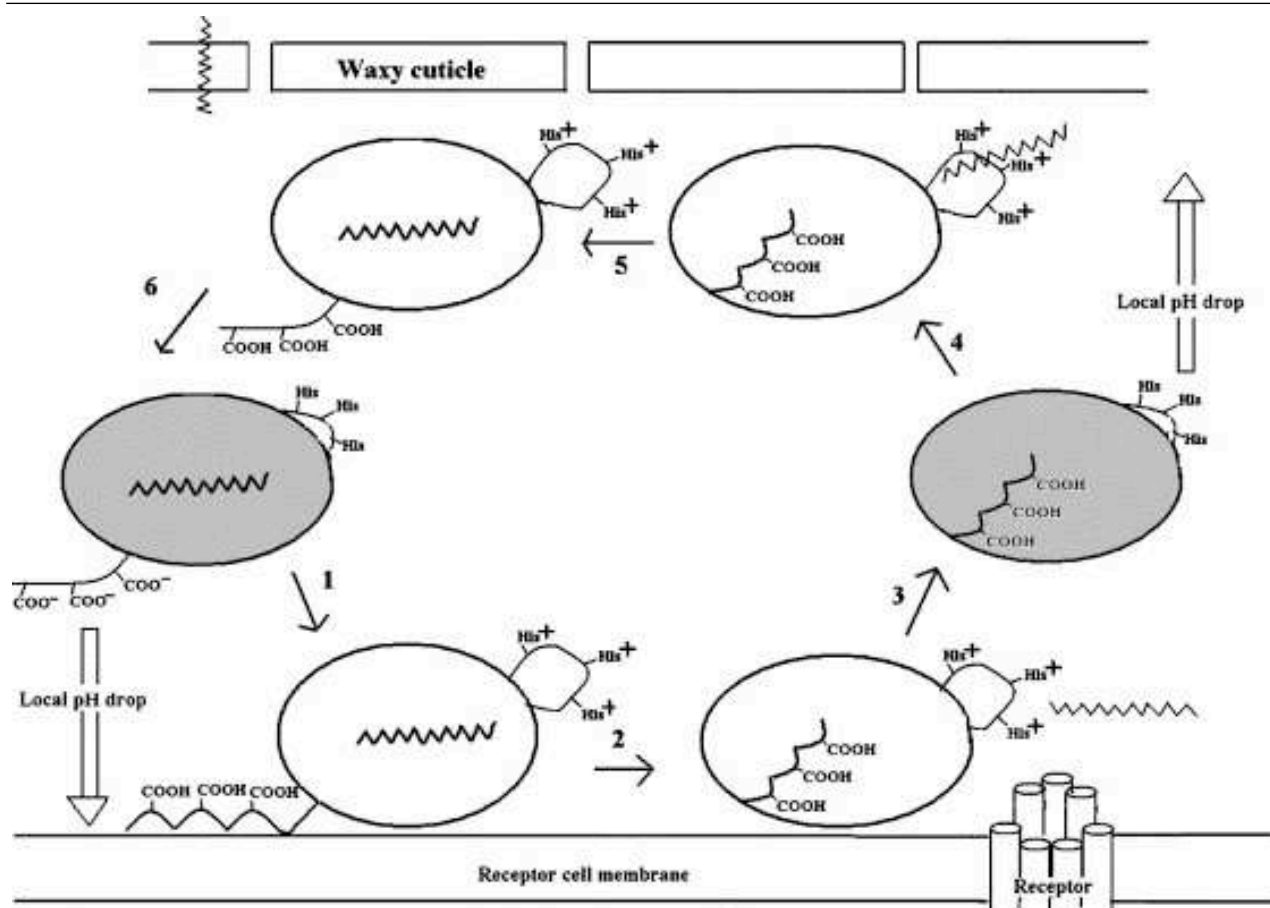


Figure 5.1. Hypothetical model for pheromone release at receptor. (1) As pheromone-bound protein approaches the membrane, C-terminal acidic residues are protonated and the C-terminus forms an ordered amphipathic helix. (2) Helix formation initiates conformational change in protein; protonation of histidine residues in loop destabilizes the region and allows pheromone to be ejected as helix $\alpha 7$ pushes into binding pocket. (3) Unliganded protein diffuses away from the membrane into higher pH region; histidine residues are deprotonated. (4) pH drop at cuticle protonates histidines in loop; loop is destabilized and ligand can enter binding pocket. (5) C-terminal tail competes with ligand for binding pocket. (6) pH increases as PBP moves away from cuticle, ionizing C-terminal acidic residues; C-terminus is no longer favored in hydrophobic-binding pocket as it is displaced by pheromone. The shaded oval represents B-form PBP; unshaded oval represents A-form PBP; pheromone is depicted as a jagged line. Histidines are indicated on looping region between helices $\alpha 3$ and $\alpha 4$; conserved acidic residues are indicated by carboxylate groups at C-terminus. Source : (Lautenschlager et al. 2005).

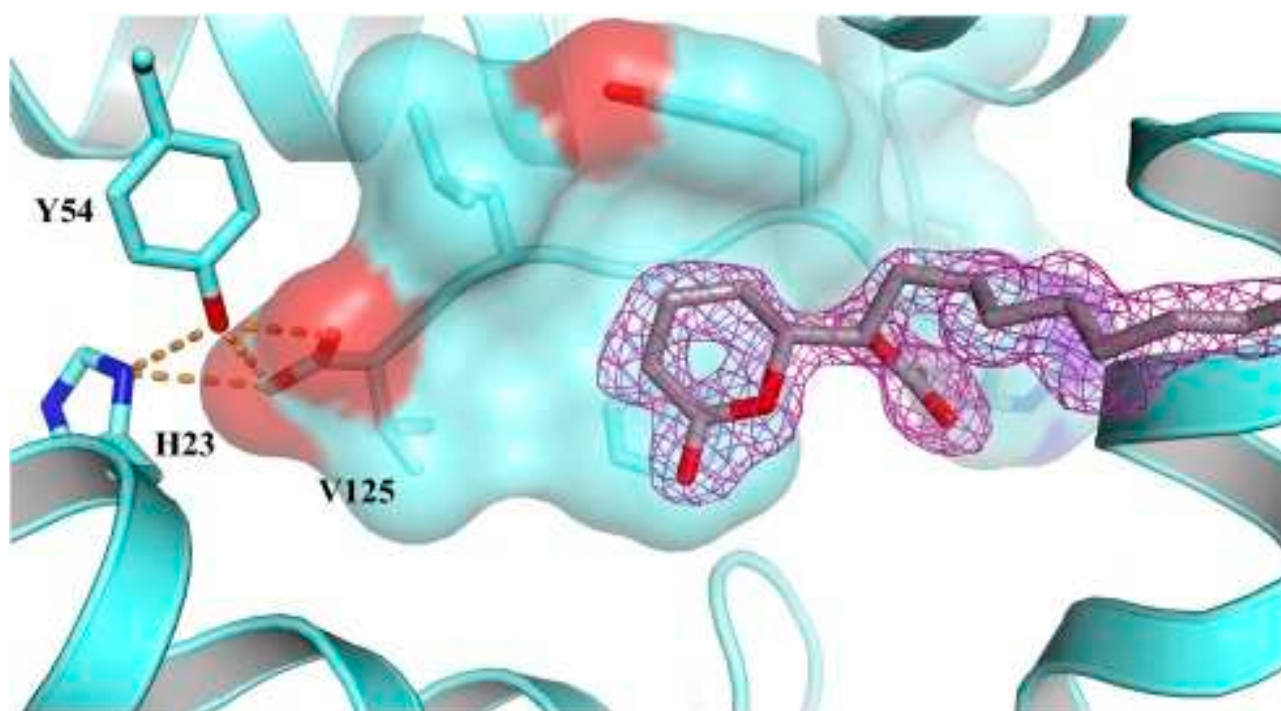


Figure 5.2. Analysis of the putative pH sensing triad between His23, Tyr54 and Val125 in CquiOBP1. Network of hydrogen bonds between His (H) 23, Tyr (Y) 54, and the C-terminal residue Val (V) 125 locks the C-terminal onto MOP, holding the pheromone molecule in the central cavity. CquiOBP1 is colored in cyan and represented in the ribbon diagram. The side chains of H23, Y54, and residues in the C-terminal are shown in stick models. A surface representation of the C-terminal of CquiOBP1 is also shown. All oxygen atoms in stick models are shown in red, and nitrogen atoms are shown in blue. Hydrogen bonds are shown as orange dotted lines. Source : (Mao et al. 2010).

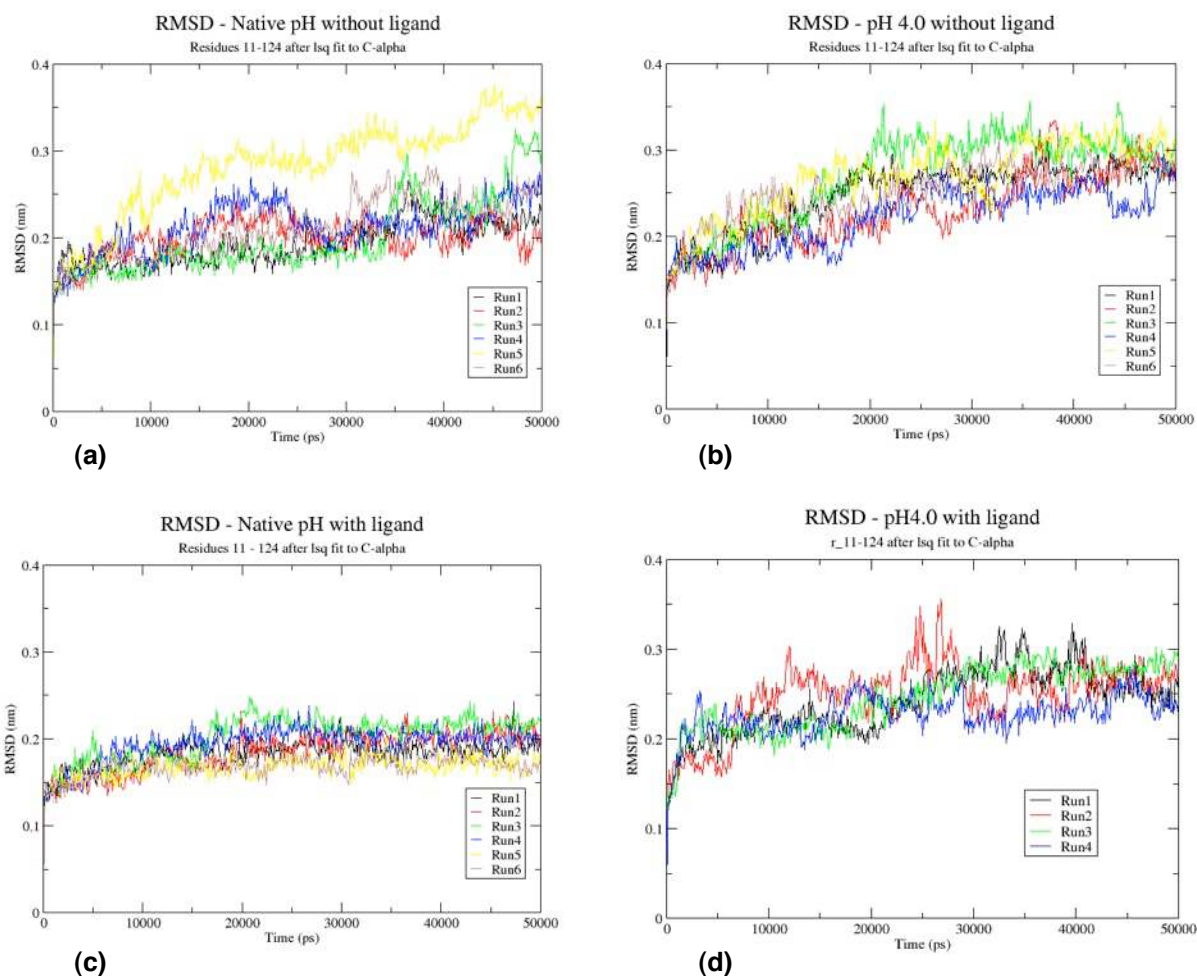


Figure 5.3. RMSD plots obtained after the least mean square fit of residues 11- 124 to C-alpha atoms of starting structure of the different simulation systems. Shown are (a) native pH without ligand, (b) pH4.0 without ligand, (c) native pH with ligand and (d) pH4.0 with ligand. The different duplicates for each simulation system are represented in different colors.

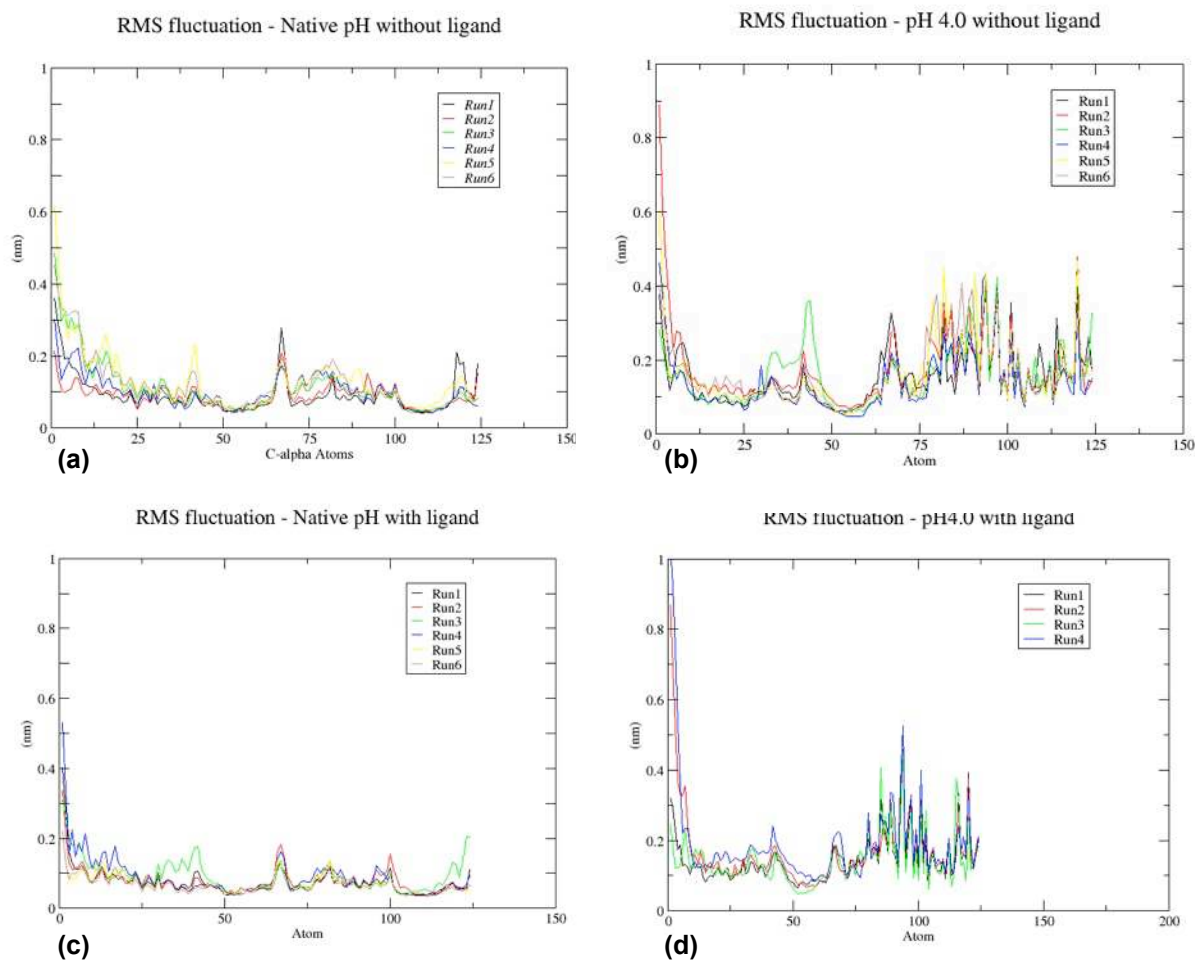


Figure 5.4. RMS fluctuation plots for C-alpha atoms of every residue in the different simulation systems : (a) native pH without ligand, (b) pH4.0 without ligand, (c) native pH with ligand and (d)pH4.0 with ligand. The different duplicates for each simulation system are represented in different colors.

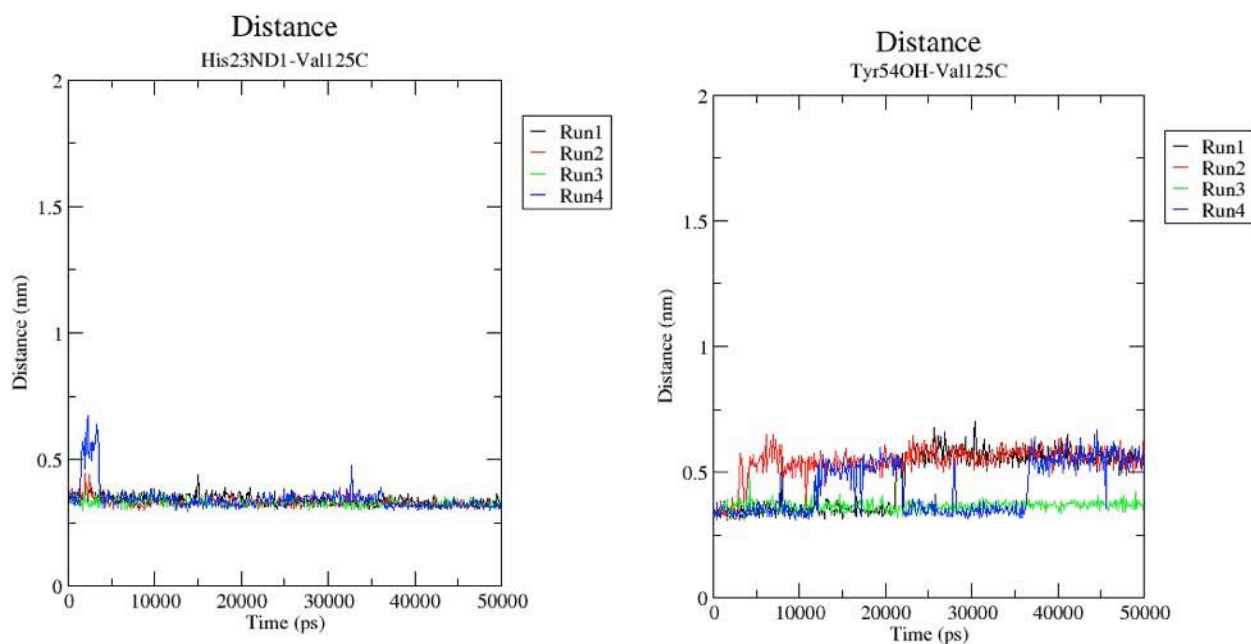


Figure 5.5. Distances for the residues involved in the hydrogen bond triad involving the C-terminus of CquiOBP1 : (a) distance between the δ -nitrogen of His23 and the carboxylate C of Val125 throughout the simulation, (b) carboxylate group of Val125 with the hydroxyl group of Tyr54 throughout the simulation.

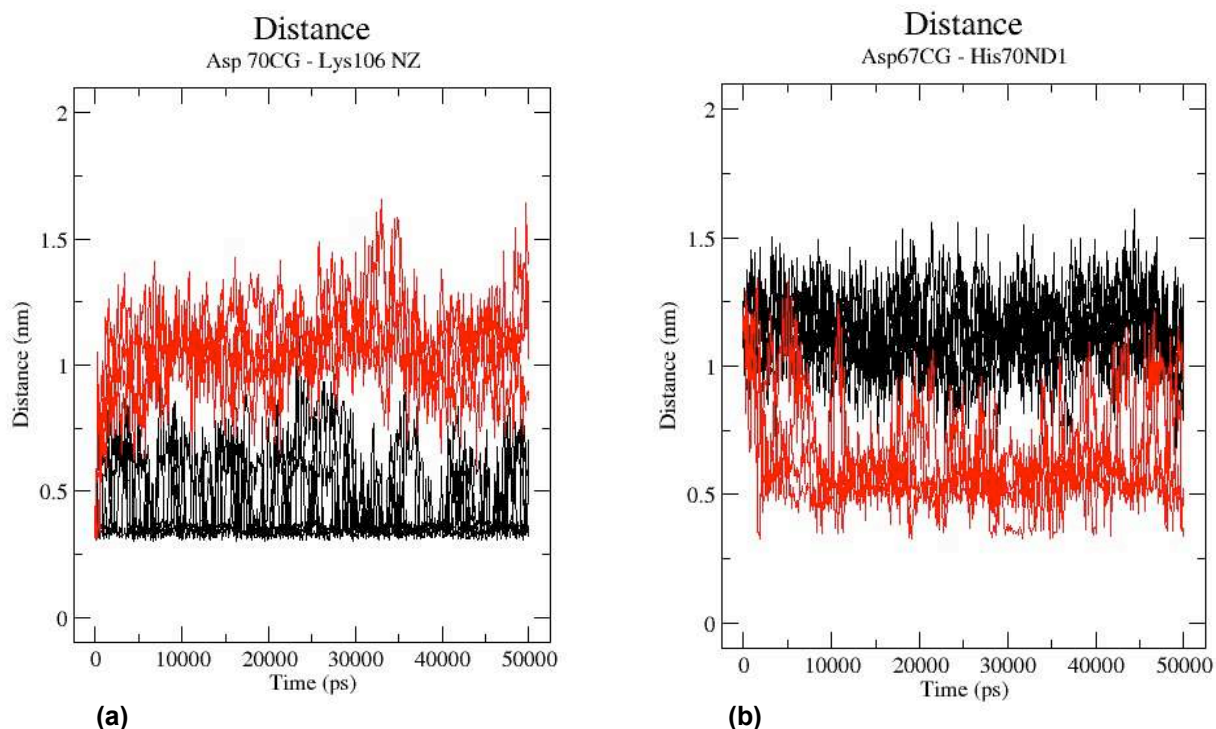


Figure 5.6. Analysis of disruption and formation of salt bridges during molecular dynamics of CquiOBP1 at pH 4.0 and pH8.0. Salt bridge involved in the change in the conformational state of the loop. (a) Shown is the disruption of a salt bridge between Asp70 and Lys106 as a measure of distances between the NZ and CG atoms of Lys106 and Asp70 throughout the simulation. The high pH simulation distances of 6 different duplicates are represented in black. The low pH simulation distances of 4 different duplicates are represented in red. (b) Shown is the formation of a new salt bridge between Asp70 and Lys106 as a measure of distances between the NZ and CG atoms of Asp67 and His60 throughout the simulation. The high pH simulation distances of 6 different duplicates are represented in black. The low pH simulation distances of 4 different duplicates are represented in red.

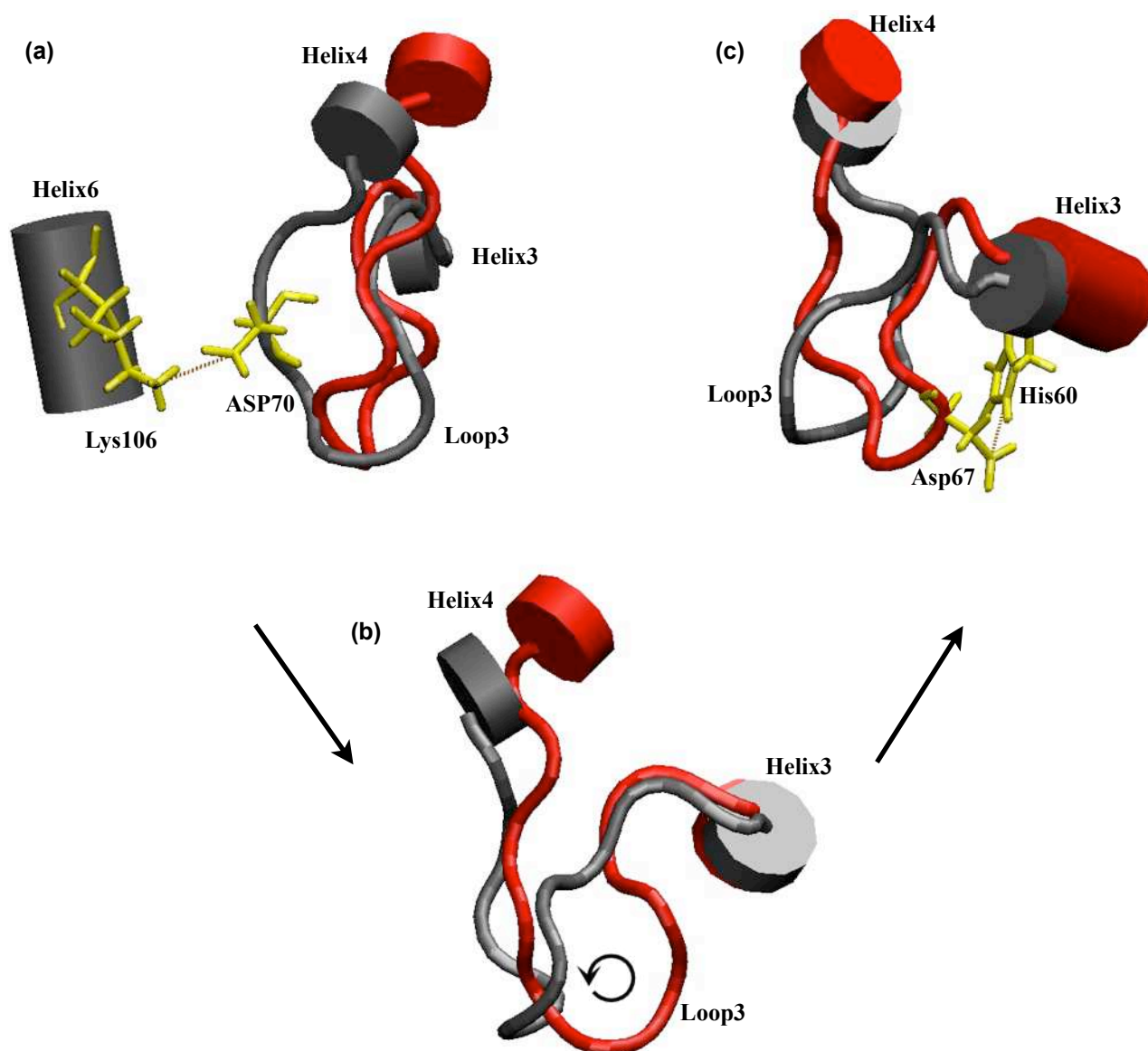


Figure 5.7. Schematic representation of the change in the conformational state of loop3 during molecular simulation at pH 4.0. Initial state of the structure is represented in gray and final state of the structure at the end of 50ns simulation is represented in red. (a) Here is featured the presence of a salt bridge between the loop3 and helix6. (b) The observed change in the conformation state of the loop. (c) Formation of a new salt bridge between the loop3 and helix3.

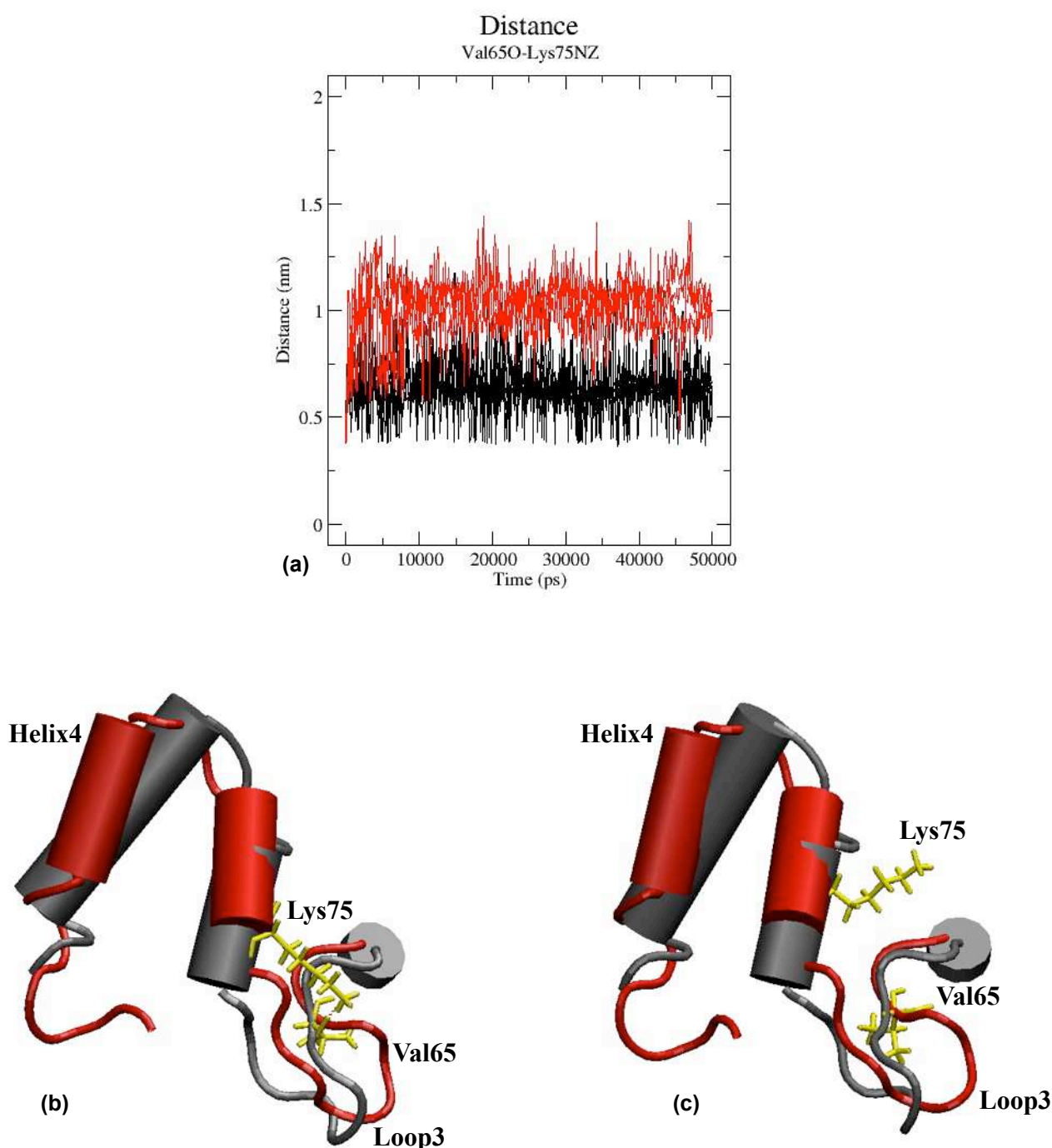


Figure 5.8. Analysis of hydrogen bond swapping associated with the change in the conformational state of the loop 3 during molecular dynamics of CquiOBP1 at pH 4.0. Initial state of the structure is represented in gray and final state of the structure at the end of 50ns simulation is represented in red. (a) Disruption of a hydrogen bond between Val65 and Lys75 as a measure of distances between the NZ and C atoms of Lys75 and Val65 throughout the simulation. The high pH simulation distances of 6 different duplicates are represented in black. The low pH simulation distances of 4 different duplicates are represented in red. (b) Presence of hydrogen bond between Val65 and Lys75 in the initial state of the structure represented in grey. (c) Disruption of hydrogen bond between Val65 and Lys75 in the initial state of the structure represented in grey and final state in red.

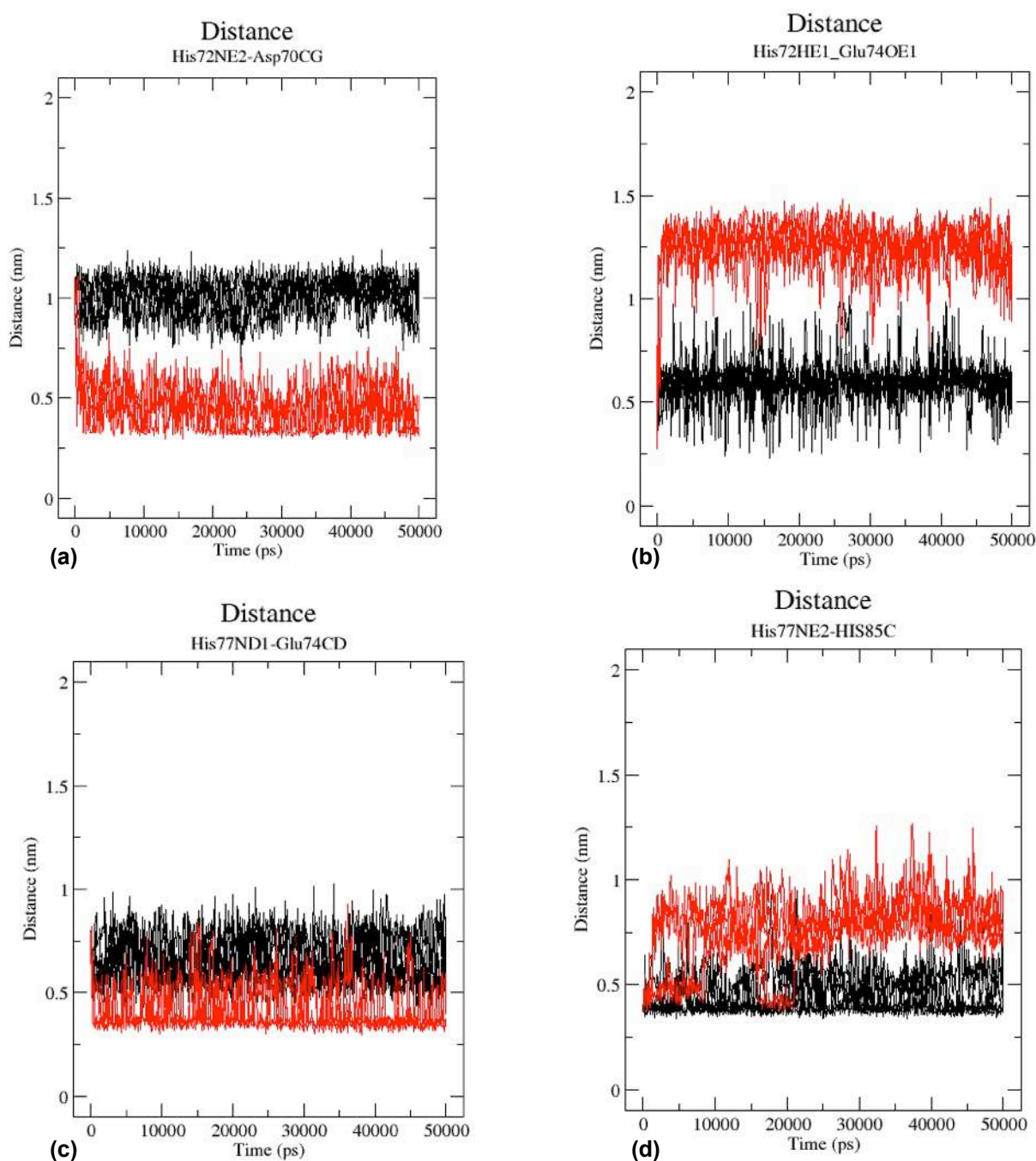


Figure 5.9. Change in the interaction pattern within and between helix4 and helix5 of CquiOBP1 during molecular simulation at pH 4.0. (a) Shown is the formation of a new interaction between His72 and Asp70 in the case of low pH simulations measured as a distance between δ -nitrogen of His72 and carboxylate carbon of Asp70 throughout the simulation. (b) Shown is the loss of interaction of Glu74 with His72 in the case of low pH simulations as a measure of distances HE1 and OE1 atoms of His72 and Glu74. (c) Is represented the new interaction between Glu74 and His77 in the case of low pH simulations as a measure of distance between ϵ -nitrogen of His 77 and the carboxylate carbon of Glu74. (d) Is represented the loss of interaction between His77 and His85 in the case of low pH simulations as a measure of distance between main-chain oxygen of His85 and ϵ -nitrogen of His 77. The high pH simulation distances of 6 different duplicates are represented in black. The low pH simulation distances of 4 different duplicates are represented in red in both the plots.

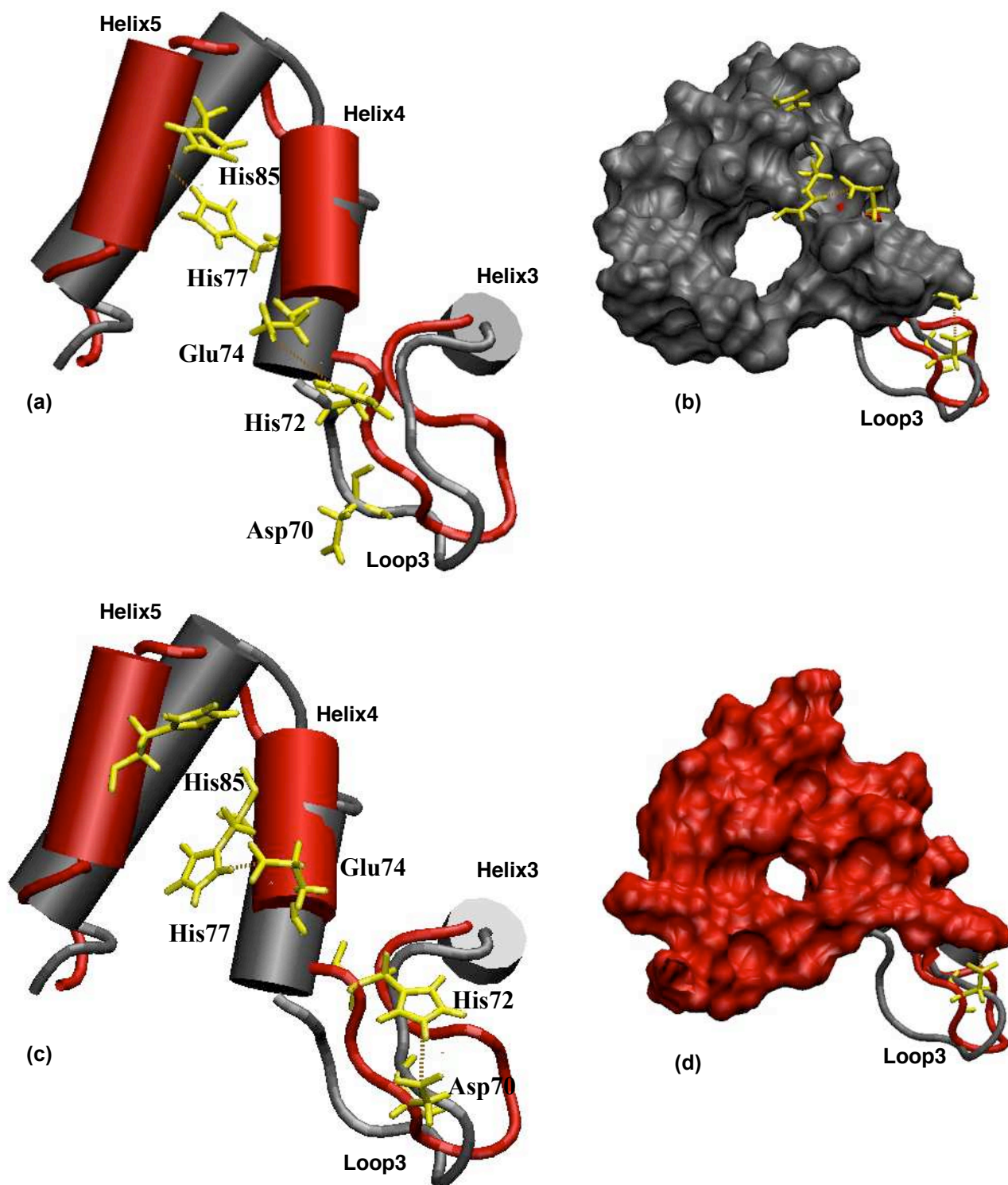


Figure 5.10. Schematic representation of the change in the interaction pattern of helix4 of CquiOBP1 during molecular simulation at pH 4.0. (a) Shown is the interaction of residues in helix4 at the initial state of the structure featured in grey depicting the interaction between His72 with Glu74 and His77 with His85 and (b) the surface of the entrance of the tunnel at the initial state. (c) Shown is the interaction of residues in helix4 at the final state of the structure at the end of 50ns simulation featured in red depicting the interaction between Asp70 with His72 and Glu74 with His77 and (d) the change in the surface of the entrance of the tunnel at the initial state due to the change in the interaction pattern.

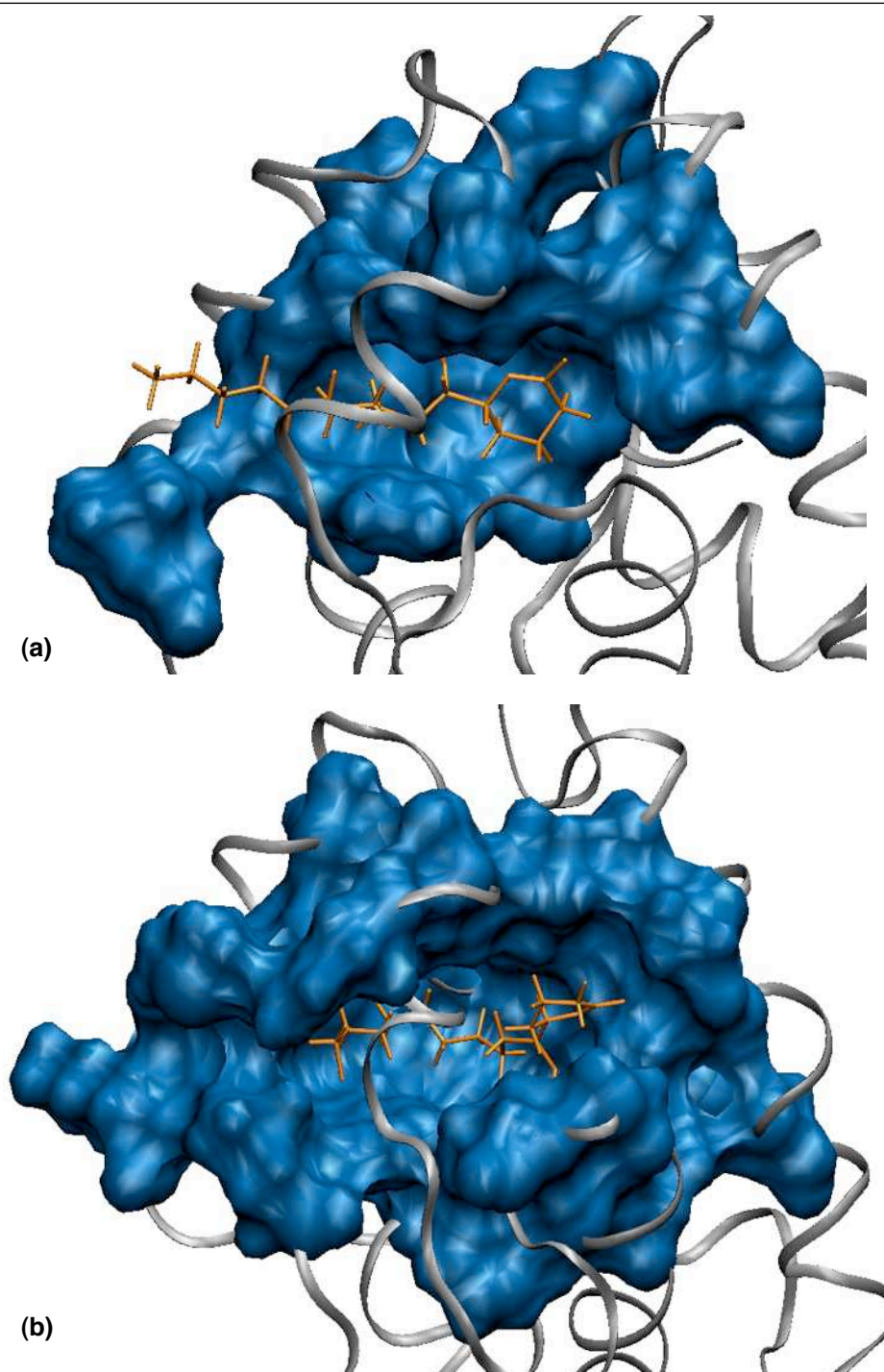
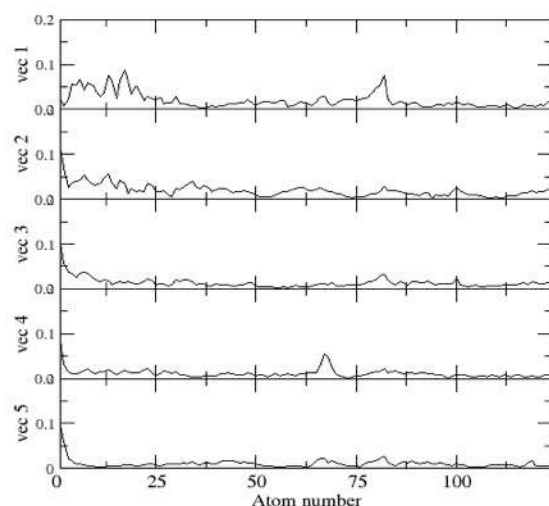


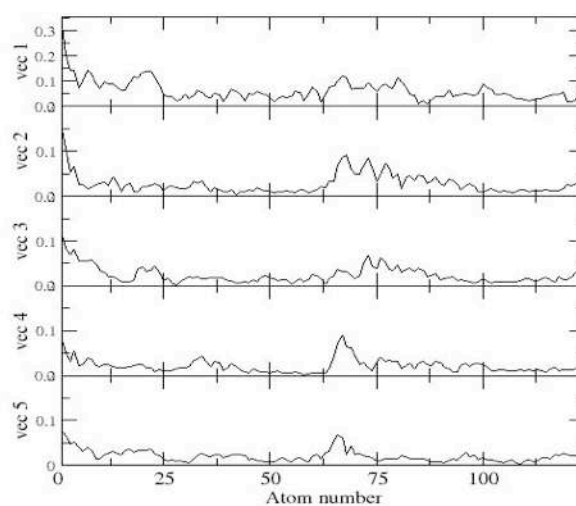
Figure 5.11. Analysis of the dynamics of the MOP ligand and its binding cavity in CquiOBP1 during molecular simulation at pH 4.0. (a) Surface representation of the residues interacting with MOP as described in the crystal structure. (b) Surface representation of the residues interacting with MOP at the end of simulation involving residues at the convergence of helix1, helix4 and loop3 featuring the extension of the tunnel ending at convergence of helix1, helix4 and loop3 - a possible exit site for the ligand.

RMS fluctuation - Native pH with ligand (nm)



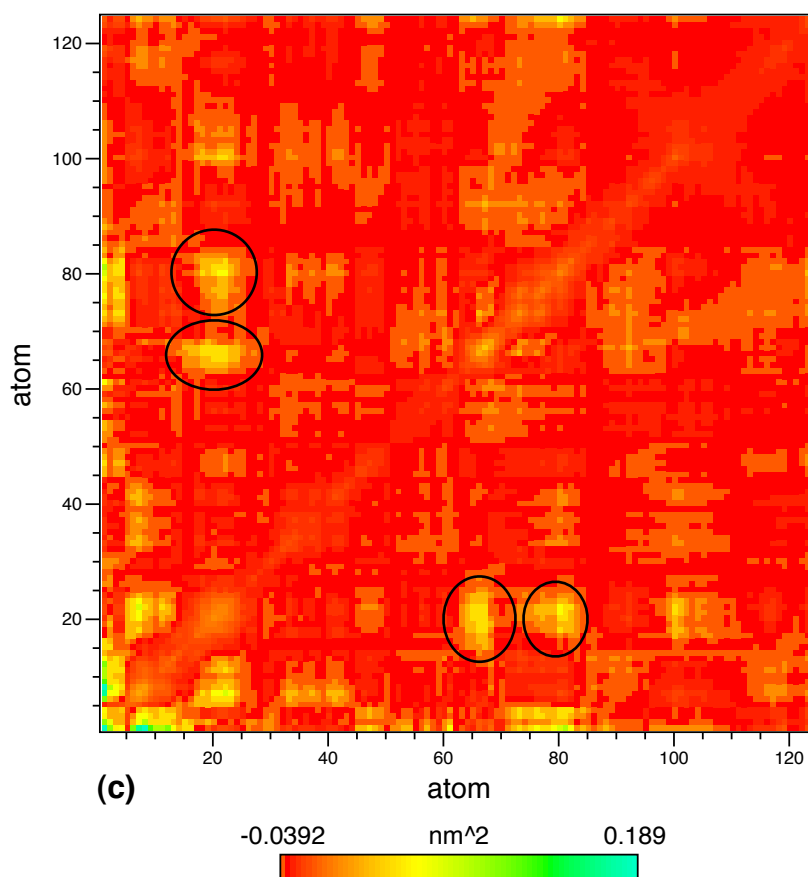
(a)

RMS fluctuation - pH4 with ligand (nm)



(b)

Covariance



(c)

Figure 5.12. Essential dynamics analysis of CquiOBP1 at pH 4.0. Covariance analysis of the movements involved in low pH simulation. (a) RMS fluctuation of C-alpha residues of the first 5 vectors at high pH simulations. (b) RMS fluctuation of C-alpha residues of the first 5 vectors at low pH simulations. (c) Covariance plot of simulations at low pH. The peaks observed at residues corresponding to loop3 and helix4 are circled.

Table 5.1. Description of the various time and pH conditions of the various simulations performed on CquiOBP1 (0PDB:3OGN) in this study.

No	Starting structure	Time	pH
MD 1 to 6	3OGN without ligand	50ns each	8.2
MD 7 to 12	3OGN with ligand	50ns each	8.2
MD 13 to 18	3OGN without ligand, protonated histidines	50ns each	4.0
MD 19 to 22	3OGN with ligand, protonated histidines	50ns each	4.0

Table 5.2. Salt bridge and hydrogen bond interactions of loop3 in the crystal structure of CquiOBP1.

Residue no	Residue	Atom name	Residue no	Residue	Atom name	Distance in Å°
63	LYS	NZ	60	HIS	O	3.11
68	ASN	N	66	ASP	OD1	2.8
69	GLY	N	66	ASP	OD1	3.1
70	ASP	N	66	ASP	OD1	2.99
106	LYS	NZ	70	ASP	OD1	2.68
75	Lys	NZ	65	VAL	O	3.21

6

Protein-ligand interaction profiles of *Classic* odorant binding proteins in the mosquito genome using molecular docking

6.1. Introduction

Olfaction is the primary cue for a number of insect species for many primary processes like food detection, host seeking, mating, oviposition and also in the identifications of predators. Unlike the mammalian olfactory system where the number of ligands recognized by the olfactory is smaller than the number of receptors present in the olfactory systems the olfactory factory systems of insects holds the reverse where the number of ligands is far more than the number of receptors in the olfactory systems. This has made the characterization of how insects handle a huge odorant space with limited number of receptors an active component of today's research. Among the insects, the olfactory system of *Drosophila* stands to be the most widely analyzed system. A combinatorial model of odor coding was established from studies on the olfactory receptors of *Drosophila* (Hallem and Carlson 2006) consistent to the generally accepted combinatorial model of OR specificity (Malnic et al. 1999). Odorant binding proteins as the primary transports of odorant molecules are expected to make significant contributions to the selectivity of the olfactory system. However given a large odorant space which is way higher than the number of OBPs, it is speculated that each OBP maybe capable of recognizing a class of structurally related odorants, and also be able to distinguish chemically different odorants. Functional dissection of odorant binding protein genes in *Drosophila* (Swarup et al. 2011) provides direct support speculating that a combinatorial activation of OBPs precedes a combinatorial activation of odorant receptors.

In the case of the mosquitoes, for *Anopheles gambiae* olfactory receptors, it was demonstrated that individual receptors respond to subsets of odorants and individual odorants activate subset of odorant receptors consistent with the above idea of a combinatorial model for

odor coding (Carey et al. 2010; Wang et al. 2010). The results were based on the analysis of a large repertoire of 72 AgOR genes against 110 odorants which generates a dataset of 5,500 odorant receptor combinations. However studies pertaining to the binding of odorants to the OBPs in the mosquito genomes is quite limited when compared to the ORs.

The perception of indole by AgamOBP1 by studying the EAG-responses (electroantennogram recording) of mosquitoes before and after silencing of AgamOBP1 in the presence of indole stands among one of the first interesting study of ligand binding in the mosquitoes. It was shown that the silencing of AgamOBP1 altered the EAG response of indole and 3-methyl indole. The *in vitro* binding of indole to AgamOBP1 showed that it had a dissociation constant $K_d = 2.3\mu\text{M}$. In the same study it was observed that the silencing of AgamOBP1 did not effect the EAG responses of mosquitoes to terpene and geranylacetone (Biessmann et al. 2010). Similar studies showed that the silencing of CquiOBP1 alters the EAG responses of MOP, Skatole and Indole suggesting that the CquiOBP1 is required to recognize these molecules (Pelletier et al. 2010). Following this (Qiao et al. 2010) studied the binding of six recombinant OBPs using a small set of organic compounds using fluorescent displacement assays. The six OBPs tested include AgamOBP1, AgamOBP3, AgamOBP4, AgamOBP12, AgamOBP19 which are *Classic* OBPs and AgamOBP47 which is a *PlusC* OBP. It was observed that these OBPs showed a broad specificity where each protein preferentially binds to several related compounds. citronellal was suggested to be the best ligand for AgamOBP1, 2 octenal and 2 nonenal for AgamOBP3 and menthol for AgamOBP4. AgamOBP12 and AgamOBP19 were found to be tuned to larger molecules where OBP12 preferentially binds to aromatic compounds and AgamOBP19 binds to terpenoids of larger size than farnesol which is its best natural ligand. The binding experiments however contradicted the previous experiments of gene silencing of AgamOBP1 and CquiOBP1 with respect to indole binding, as they did not find any of the OBPs that showed significant affinity to indole.

The crystal structure of CquiOBP1 complexed with MOP, an oviposition pheromone, is the first ligand bound structure of a mosquito OBP which provides a very good insight into the binding pocket of the mosquito OBPs. The ligand binding pocket of CquiOBP1 similar to the other OBP structures is found to be a central cavity inside the protein covered by hydrophobic residues. In addition a major part of the the ligand was bound to a hydrophobic tunnel formed between helix4 and 5. Separately the binding assays in this study showed that Octanal, nonanal, decanal and geranylacetone showed significant binding affinity to CquiOBP1 while γ -Octalactone did not show any binding affinity (Mao et al. 2010). More recently the crystal structure of DEET bound to AgamOBP1 was described by (Tsitsanou et al. 2011). DEET was also found to bind to the edge of a

long hydrophobic tunnel as described previously in CquiOBP1. This study also reports *in silico* studies of AgamOBP1 binding to a few potential insect repellants using Autodock and found that few molecules showed higher binding efficiency to DEET. The validity of the modeling predictions were tested experimentally and it was found that the testing compounds indeed showed increase capacity of the ligands to bind to AgamOBP1. And a fairly good correlation of the experimental K_d values and Autodock K_i values were observed indicating that docking simulations can be a good starting point for experimenting a large number of compounds. Previous studies on the OBPs of mosquito genomes have focussed on one or a small number of receptors and ligands. However a global perspective in which a large screening of the family over a wide odorant space is required to provide a better understanding of the functional repertoire of the OBPs in the mosquito genomes. This study provides the first insight to the function of OBPs based a large scale screening of 125 odorants against 129 classic OBPs in the mosquito genomes using computational approaches.

6.2. Materials and methods

6.2.1. Construction of the ligand database

110 ligands used in the analysis were derived from the previous dataset of odorants used in the functional characterization of odorant receptors in *Anopheles gambiae* (Carey et al. 2010). 53 of these odorant molecules were also tested on the *Drosophila* antennal receptor repertoire (Hallem and Carlson 2006). They constitute a chemically diverse set of compounds including acids, ketones, alcohols, terpenes, esters, amines, aldehydes, aromatic and heterocyclic compounds of various molecular sizes (Table 6.1). The dataset includes compounds described as oviposition site volatiles, active components of human emanations and compounds reported to change the behavioral activity of mosquitoes (Carey et al. 2010). Additional compounds which were studied to have repellant properties towards mosquitoes were also included in the current analysis (Keisuke Watanabe et al. 1993; Fradin and Day 2002; Barnard 2005). The 3D coordinates of these compounds were downloaded from the Pubchem database (<http://pubchem.ncbi.nlm.nih.gov>) of chemical compounds and substances. The downloaded ligands were then converted into a AUTODOCK accepted format for further use.

6.2.2. Docking

The molecular docking program AUTODOCK 4.0 was used to dock the various ligands to the different receptors. Prior to the large scale docking various combination of parameters in

AUTODOCK were tested on the known protein ligand complex of CquiOBP1-MOP complex. It was the only crystal structure that was available when this study was started. The unbound protein was energy minimized before using with the docking program to remove any short contacts caused by the extraction of the ligand. The minimization was done in a water box with OPLS-AA forcefield parameters using Gromacs 4.5.3. The energy minimized structure file was further prepared as required by the AUTODOCK program by adding hydrogens and gasteiger charges using ADT tools. The 3D coordinates of the ligand MOP was obtained from the Pubchem database. The detected 12 rotatable bonds in the ligand were chose to be retained. Lamarkian genetic algorithm search method with different combinations of grid parameters for Autogrid and dock parameters for AUTODOCK were tested (Table 6.2) in order to optimize the parameters which could reproduce a conformation close to the determined crystal structure. The results (see below) indicated that the 9th set of parameters in the Table 6.2 could efficiently reproduce the position and conformation of the ligand. Consequently a large scale docking was carried out using a box size of 70*54*66 with a grid centre of (31,156; 38,673; 38.01) which closely projects the binding pocket of the OBPs. Lamarackian genetic algorithm was used with 25,00,000 evaluations with 100 GA runs with a population size of 150 and all the ligands were treated as flexible. An parallel-version of the program developed by (Khodade et al. 2007) which can be used on cluster machines was used for the analysis. The docking was carried out using 48 processors on the cluster facility in the University of de La Reunion. The total computation time was reduced from 398 days (on a single node machine) to 9 days with the efficiency of the cluster. The docking was carried out in a automated fashion for the huge dataset. The results were further analyzed using python scripts available with MGLtools and other in house scripts.

6.2.3. Estimation of significant interactions

Several computational approaches used for assessing the binding of the ligand to a protein generally provide the free energy of binding value which is unfortunately directly dependent to the size of the ligand. Optimizing the binding potency of a ligand independent of the size of the ligand has been widespread over the community in order to improve large scale screening experiments in drug discovery programs. One such measure is called the Ligand efficiency (LE) which is calculated using the formula described in equation (1). However it was observed that the ligand efficiencies show a marked decrease with increasing molecular size, especially for ligand with less than 20 atoms.

$$LE = \frac{affinity}{HA} \quad (1)$$

Therefore another metric was developed called the fit quality score (FQ) as a direct measure of how optimally a ligand binds relative to other ligands of any size. It is calculated using the formula in equation (3) where LE_Scale is a scaling parameter derived from fitting ligand efficiency values for ligands with 10-50 atoms as shown in equation (2). It was suggested that a calculated LE_Scale value at HA=15 to be taken for any compound with 15 or fewer.

$$FQ = LE / LE_SCALE \quad (2)$$

$$LE_Scale = 0.0715 + 7.5328 / (HA) + 25.7079 / (HA^2) - 361.4722 / (HA^3) \quad (3)$$

Subsequently another size independent measure was suggested called Size independent ligand efficiency measure (SILE) which is calculated as in equation 4.(Nissink 2009) where affinity may be either free energy of binding or pKi values. Here, it is the pKi values that was used.

$$SILE = \frac{affinity}{N^{0.3}} \quad (4)$$

In the analysis the number of HA atoms was calculated for each of the ligand in the ligand dataset. The lowest binding energy of each of the docked complex was converted to Ki values as in equation 5.

$$Ki = \exp \frac{\Delta G * 1000}{RT} \quad (5)$$

$$R = 1.98719(cal) \quad T = 298.15(kelvin)$$

The corresponding LE, FQ and SILE values were calculated for all the complexes. All the plots used in the analysis of the results were generated based on these values using the R software.

6.3. Results

6.3.1. Optimization and validation of docking protocol

The bound structure of CquiOBP1-MOP crystallized complex was used in the evaluation of the docking program and to optimize the various parameters of AUTODOCK for the large scale docking experiment on the Classic OBPs. The long lipid tail of MOP was found to be bound to the hydrophobic tunnel formed between helix4 and helix5 and the lactone/acetyl ester occupies the

central cavity involving residues Tyr10, Leu19, Leu80, Ala88, Met91, His111, His121, Tyr122 and Phe 123 in the crystal structure. From the different parameters tested, the parameters highlighted in bold in Table 3.2 were found to be optimal for reproducing the binding of MOP to CquiOBP1. However the orientation of the carboxylate oxygen of the lactone ring was different from the docked complex (Figure 6.1). The residues that were found in close contact with the ligand are Tyr10, Leu19, Leu58, Leu73, Leu76, His77, Leu80, Ala88, Met89, Gly92, His111, Tyr122, Phe123, Leu124, Val125. The residues found in close contact with the ligand in both the crystal structure and the docked complex are quite similar.

6.3.2. Docking

Docking experiments were carried out on 129 proteins against 125 ligands as described in the materials and methods. A total of 16,380 lowest binding energy for each of the ligand against each protein model was extracted in the form of a table from the output files and used for further analysis. All the 100 conformation generated from every docking experiment have been reproduced as a docked complex making a total of 1,638,000 PDB files which can be accessed on our website. The table of the energies has been reproduced as a heat map for visualization purpose (Figure 6.2).

6.3.3. Analysis of the binding efficiency

It is very well known that the AUTODOCK binding energy is directly proportional to the ligand size and thus it does not allow a fair comparison of the data. In order to further analyze the data it is important that a size independent measure is used to test the efficiency of the binding and allows for an agreeable comparison of the data. The table of the docking energy was thus converted to a table containing K_i values from which the LE, FQ and SILE values were calculated as described in materials and method. It was observed that when the free binding energy was plotted against the number of heavy atoms (HA), the energy value increased with the atom size (Figure 6.3a). A similar trend was seen for the K_i values also (Figure 6.3b). The plot of LE against the number of heavy atoms suggested that smaller ligands are more efficient binders but are still not a good measure as the ligand efficiency shows a gradual decrease as the size of the molecule increases (Figure 6.3c). The corrected Ligand Efficiency also indicated a size dependency (Figure 6.3d). However the SILE measure suggested by Willem et al. (2009) was found to be a size independent measure for the current dataset as shown in Figure 6.3e. This measure was used for filtering out significant interactions. The values of the SILE measure ranged from 1.0 to 3.5 (Figure

6.3f). The highest population of data was found to have an SILE value between 1.5 to 2.0. The minimum SILE value observed is 0.88, the first quartile value is 1.54, the second quartile/median value is 1.73, third quartile is 1.94 and the maximum value is 3.71. In order to set a threshold value for identifying the significant interactions, the SILE values of some of the ligands which are experimentally showed to bind to the proteins were considered (Table 3.3). Based on the observed patterns a threshold of 1.92 was set as the upper threshold and a lower threshold of 1.66 was set up in general. However these thresholds could still vary for some ligands but we still used them as there is a lack of other experimental support. A heat map of SILE values was generated and it can be seen that it gives a more clear and unbiased representation of the data when compared to the heat map generated using free energy of binding (Figure 6.4).

6.3.4. General overview on binding

The SILE values of each of the ligand which is directly proportional to the efficiency of binding were plotted with the receptors on the X-axis with the highest SILE value in the middle (supplementary material 3a). The kurtosis value was calculated and is indicated on the graph. These ligand tuning curves give a good representation of the ligands which bind specifically to the receptors and the ligands which bind to a large number of odorants. If the upper threshold obtained from experimental data is applied on these tuning curves we found that some ligands did not show significant binding to any of the receptors. Odorants like 1-butanol, 2,3 butanedione, 2-butanone, dimethylsulphide, ethanol, ethyl acetate, ethyl formate, ethyl propanoate, methanol, methyl propionate, thiazole, 2ethoxythiazole, propyl acetate, acetone do not show high binding efficiency to any of the receptors (Figure 6.5). Some of the ligands were found to efficiently bind to a large number of receptors (curves with a low kurtosis value which included the terpenes, 5- α -androst-16-en-3 α -ol, 5- α -androst-16-one, E2-hexenal, PMD, ammonia, cadaverine, linalool oxide, permethrin, putrescine (Figure 6.6). In contrast it was observed that some of the ligands exhibited high kurtosis values indicating that bind quite specifically to the receptors. 2-ethyl toluene, 2-ethylphenol, 2-methylphenol, indole, 3-methylindole, 3-methylphenol, 4-ethylphenol, 4-methylcyclohexanol, 4-methylphenol, Nepetalactone, benzaldehyde, phenol, acetophenone (Figure 6.7). All the repellants used in the dataset showed a low kurtosis value indicating a broad specificity for the receptors.

6.3.5. OBP binding profiles

If we were to make similar plots with respect to the OBPs they would give an idea about the receptors that are narrowly tuned and the ones that are more generalists binding to a larger set of ligands (supplementary material 3b). In general a broad specificity of the OBPs to bind to different ligands is observed. The results indicated that the OBPs belonging to the Pbprp2/Pbprp5 cluster in *Culex*, members of Lush, Pbprp1, *mclassic7*, *mclassic8* and OBP members closely related to *Bombyx mori* minus C proteins and *Drosophila* minus C proteins showed high kurtosis value indicating that they could be more tuned to specific odorant molecules. Members of the other clusters show a broad specificity to the ligands. A box plot was constructed for each of the ligands grouping all the members in a cluster (supplementary material 3c). The box plot gives a representation of the variation of binding efficiency within the cluster and also can give the specificity of a cluster as a whole to bind to particular odorant molecules. The results indicated that the OBPs belonging to MClassic2 cluster always showed a high binding efficiency broadly to all the acids while the MClassic7 always indicated a low binding efficiency for acids. All the repellants used in the ligand dataset showed significant binding efficiency among the clusters however the highest binding efficiency for all the clusters was observed for permethrin. Highest binding efficiency for all the clusters is observed for ammonia and 5- α -androst-16-en-3 α -ol, 5- α -androst-16-one (Figure 6.8). It was interesting to note that the finely tuned receptors showed high binding efficiency to broadly tuned odorants. Among them permethrin showed high binding efficiency to OBPs belonging to all the clusters (Figure 6.9).

6.3.6. Comparison of the computational docking complexes with experimental docking complexes

Subsequently after the docking experiments were completed a crystal structure of AgamOBP1 with DEET was published. On comparing the docked complex from the AUTODOCK with the crystal structure, to our surprise it was found that the conformation and binding of DEET predicted by AUTODOCK was very close to the crystal structure that was published (Figure 6.8). The DEET binding pocket was described to be formed by the residues in helix4 (Leu73, Leu76, His77, Leu80), helix5(Ala88, Met89, Met91, Gly92), helix6 (Trp144) and Leu96', Lys93', Arg94' and Leu96' where the (') refers to the dimer. The involvement of a water molecule in the binding site is also described. The results from AUTODOCK of DEET docked to AgamOBP1 showed a similar ligand conformation (Figure 6.10). However the involvement of residues from the other monomer and the water molecule could not be compared as the docking protocol was restricted to

monomers and water molecules were not included. Another crystal structure of AgamOBP4 with indole at pH 6.5 was also published immediately after the crystal structure of AgamOBP1-DEET complex was released. The ligand in the case of the docked complex was found to be located in the site located between helix3 and helix4 similar to the location of MOP in CquiOBP1 structure. However in the case of the crystal structure PDBID:3Q8i it is found situated close to the helix3 (Figure 6.11) which was previously described as the possible exit site for the ligand at low pH conditions in Chapter 5.

6.3.7. Characterization of binding site for known experimentally proven ligands of mosquito OBPs

Some of the ligands chosen in our study have been experimentally shown to bind to OBPs in the mosquito genome. The probable binding site for these ligands in the respective OBPs are shown in (Figure 6.12). Ligands which bind to AgamOBP1, AaegOBP1, CquiOBP1 which are orthologous to each other interact with more or less the same residues. Octanal, Nonanal, and Decanal interact with these OBPs by forming a hydrogen bond to Phe123 which is conserved in all the three OBPs. Geranyl acetone occupies a similar binding site in all the three receptors. Indole and 3-methyl indole however interact with different residues in all the three OBPs but they in general involve residues from helix4, helix5, helix6 and the C-terminal loop. Octanal Nonanal, Decanal and 1-dodecanol bind to AgamOBP4 involving similar binding residues. They interact with AamOBP4 by forming a hydrogen bond between Ser10 and Met77. Citronellal in contrast forms a hydrogen bond with Thr70 of AgamOBP4 and involves a slightly different set of binding residues when compared to the other ligands listed above. Nonanal and citronellal bind to AgamOBP3 with residues located in the C-terminal loop, helix3, helix4, helix5 and helix6. Both the ligands form a hydrogen bond to Phe122 on the C-terminal loop. Citronellal binds to AgamOBP19 involving residues located in helix3, helix4, helix5 and helix6.

6.4. Discussion

6.4.1. Optimization and validation of docking protocol

The docking process generally involves the prediction of the ligand conformation and orientation within a targeted binding site. The choice and preparation of the structural model of a targeted binding site are important variables in this process. It is important to consider how a protein and ligand are represented. AUTODOCK uses a grid representation for energy calculations

where the basic idea is store information about the receptors energetic contributions on grid points so that it only needs to be read during ligand scoring. It is therefore important to define these points for each protein most preferably based on previously established experimental data which can be extended to a family of proteins. It is also observed that other parameters of the AUTODOCK program such as the number of GA runs or the number of evaluations can also affect the quality of the docking. Therefore an optimization of an exact representation of the receptor and other parameters stand a important starting point for any docking experiment. Optimization in the current analysis was based on the ability of AUTODOCK to reproduce the experimentally determined docked complex of CquiOBP1-3OG complex. The various grid parameters and run parameters tested revealed a single set of parameters to be the most optimal for the odorant binding proteins. The finalized set of parameters were able to closely reproduce the ligand binding site and conformation as found in the crystal structure. The only difference that was observed was on the orientation of the lactone ring. It is to be noted that the electron density for this part of the ligand was not well defined suggesting that this part of MOP can have several conformations in the cavity. This given set of parameters were extended to the other members in this family believing on the fact that they share a similar binding cavity. However it is to noted that a fairly large grid was used and the binding was not guided or biased.

6.4.2. SILE is a good measure for a size independent representation of the data

In general, the scores of any of the docking methods, regardless of the functions applied are known to scale poorly with molecular mass and the number of rotatable bonds. Large molecules can form many hypothetical interactions in the binding sites and therefore have the tendency to generate better scores than smaller compounds which can also be seen in the current analysis. The initial question of the large scale binding studies being oriented towards understanding the specificity of ligands and receptors however requires a measure that is independent of the size of the molecule. The need of such a measure is also highly prevailed in the virtual screening community. Therefore size independent measures have been developed such as LE, FQ, SILE and they have been implemented in the current analysis. The plots of each of the measures against the number of heavy atoms in the ligand which represents the size of the ligand shows the trend observed for these measures (Figure 6.3e). The SILE measure proves to be the best size independent measure as observed for the current dataset. The use of this measure in protein-ligand docking experiments has also been previously emphasized. The FQ measure scales very badly for ligands with low number

of heavy atoms as shown here (Figure 6.3d). The parameters described (equation 2) were optimized for ligands with $HA \geq 10$ which is not suited for our dataset.

6.4.3. Variations observed in the binding profile of odorants to ORs and OBPs - suggesting the importance of the role of OBPs for certain ligands

The general ligand profiles indicate that some of the odorants are finely tuned and some show a more general profile with the ability to bind to a large number of OBPs. It was observed that some of the ligands did not show high binding efficiency to any of the receptors. Among these, ligands like 1-butanol, 2,3 butanedione, 2-butanone, dimethylsulphide, ethyl acetate, ethyl formate, propyl acetate, ethyl propanoate, methyl propionate, thiazole, 2-ethoxythiazole, which however directly and strongly activated the olfactory receptors (Carey et al. 2010) suggesting that they may act directly on the olfactory receptors and may not require the involvement of odorant binding proteins. In contrast, it was observed that some of the ligands that showed high binding affinities for OBPs did not show a direct high firing patterns for the olfactory receptors in the same study (Carey et al. 2010). Highest binding efficiency for almost all the receptors was observed for 5- α -androst-16-en-3 α -ol, 5- α -androst-16-one and ammonia. They are reported to be components of human emanations (Zeng et al. 1991; Brooksbank et al. 1974; Ellin. 1974; Braks. 2001; Czarnowski. 1991). A notable high binding profile for these ligands to all the OBPs raises the question if it could be important components involved in the host seeking mechanism. However these compounds were not found to directly activate any of the olfactory receptors with a high firing rate hence kindling the role of the odorant binding proteins. It may be that the hypothesis where the conformational change in the OBP is directly recognized by the receptors without the direct involvement of the ligand could be prevalent in this case.

A broad specificity observed for the repellent molecules is an interesting observation in this analysis and permethrin among all the other repellent molecules showed the highest efficiency to almost all the receptor indicating it could be one of the most efficient repellent molecules on the market

6.4.4. Combinatorial binding profiles observed for OBPs

Overall the odorant binding proteins in the mosquito genome show a combinatorial mode of binding to the odorant molecules. It was observed that some of the OBPs showed significant binding efficiency to a large number of odorants. It was also observed that some of the receptors in well known clusters are fine tuned to a subset of odorants. Members of the certain clusters

discussed in detail in Chapter 2 showing high kurtosis value is an interesting observation of this analysis. The members of the Pbprp2/Pbprp5 which showed an explosion in the number of members in the case of *Culex* when compared to *Anopheles* and *Aedes* show high specificity in binding to ligands indicating a functional requirement for the gene duplication observed for *Culex* for this cluster. It was also interesting to observe that certain cluster showed high ligand binding specificity (MClassic2) to certain set of ligands (acids) and another cluster showed a low binding specificity to the same set of ligands (MClassic 7). In the current analysis from the box plots that were generated it is clearly evident the MClassic2 is more tuned to bind to acids while the MClassic7 always indicated a low binding efficiency to the acids. This stands a clear representation of the differences in the binding site that is observed in par with the high sequence diversity that is observed in this interesting protein family.

6.4.5. Comparison of the docked complexes to experimental data

The ability of AUTODOCK to reproduce results close to the experimental data has been observed in many cases. It was previously observed in our analysis that given a set of fine tuned parameters AUTODOCK could efficiently reproduce the docked conformation of a ligand. Surprisingly and promisingly high correlation was also observed for the results obtained in this analysis with crystallized complexes that were published after the analysis were carried out. This indicates that the parameters of AUTODOCK used in this analysis is capable of producing considerably accurate results in the case of OBP from insects. It also indicates that AUTODOCK is a good software that can be used in large scale screening experiments for OBP in insects. The ligands that were shown to bind to the OBPs experimentally indicated a good binding efficiency towards the receptors (SILE). Characterization of the binding site based on the analysis of certain bound complexes for the ligands which have been shown to bind experimentally to OBPs highlights the importance of the residues in the C-terminal loop described to form a wall of the binding pocket in the case of the Mosquito OBPs. It was observed that most of the ligands that were characterized for the analysis of the binding site indicate the involvement of residues in the C-terminal loop for their binding. The binding site also involved residues from helix3, helix4, helix5 and helix6.

6.5. Conclusion

The current analysis not only provides a global picture on the functional repertoire of odorant binding proteins but also raises many important biological hypothesis and provides support to some of the the previously prevailing hypothesis on ligand binding mechanism of the OBPs. The OBPs of mosquito in general follow the previous established combinatorial model of odor coding in the insect species where individual receptors respond to subsets of odorants and individual odorants activate subset of odorant receptors. Though the accuracy of these results observes various checkpoints throughout the entire procedure it still stands to be a valuable analysis giving primary insights into the functional repertoire of the OBPs in mosquitoes.

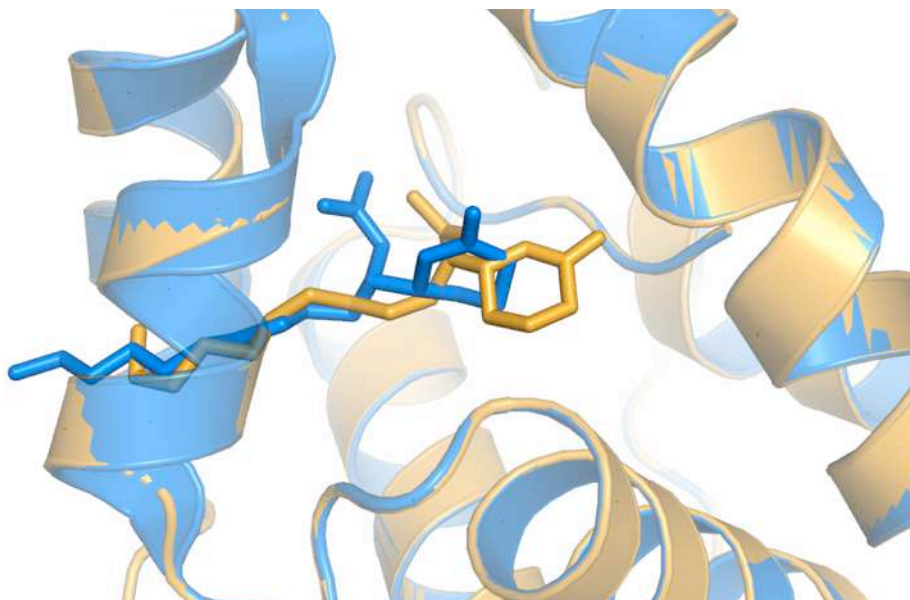


Figure 6.1. Structural superposition of MOP bound to AgamOBP1 as in the crystal structure and the predicted complex. The ligand in the crystal structure is represented in blue and in the docked complex is represented in orange.

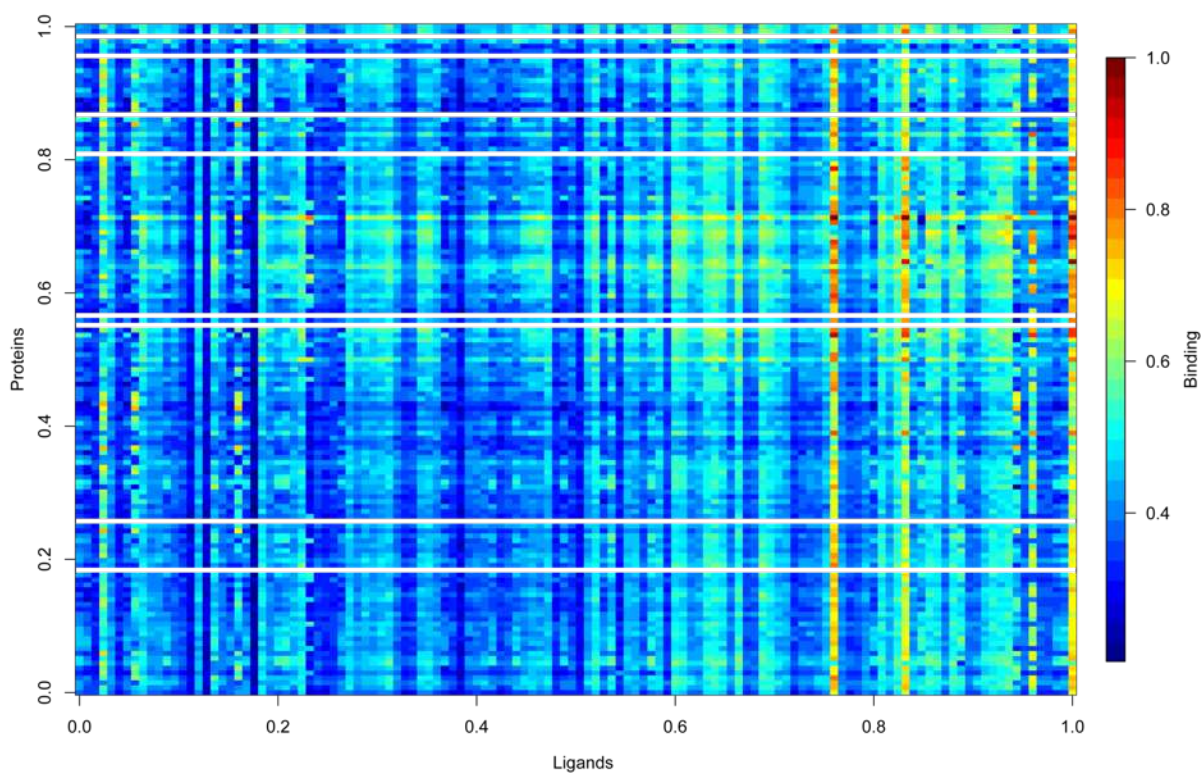


Figure 6.2. The docking free energies of 130 proteins against 126 ligands represented as a heat map.

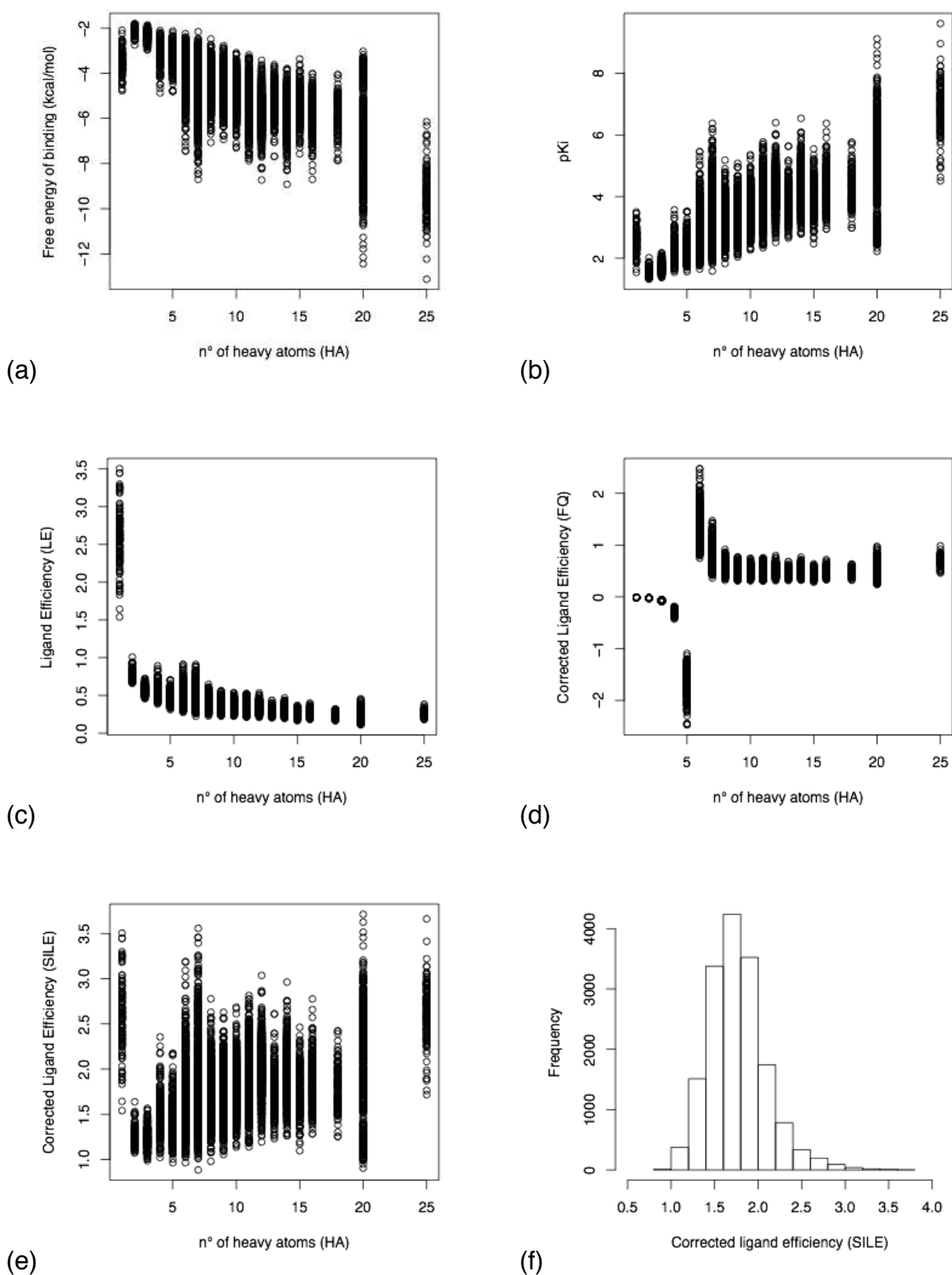


Figure 6.3. Various plots constructed on arriving at a size independent measure of the data. (a) Plot of free energy of binding (FEB) against HA atoms of the ligand, (b) plot of pKi vs HA atoms, (c) plot of ligand efficiency(LE) vs HA atoms, (d) plot of corrected ligand efficiency (FQ) against HA atoms, (e) plot of size independent ligand efficiency (SILE) against HA atoms and (e) distribution of SILE values in the dataset.

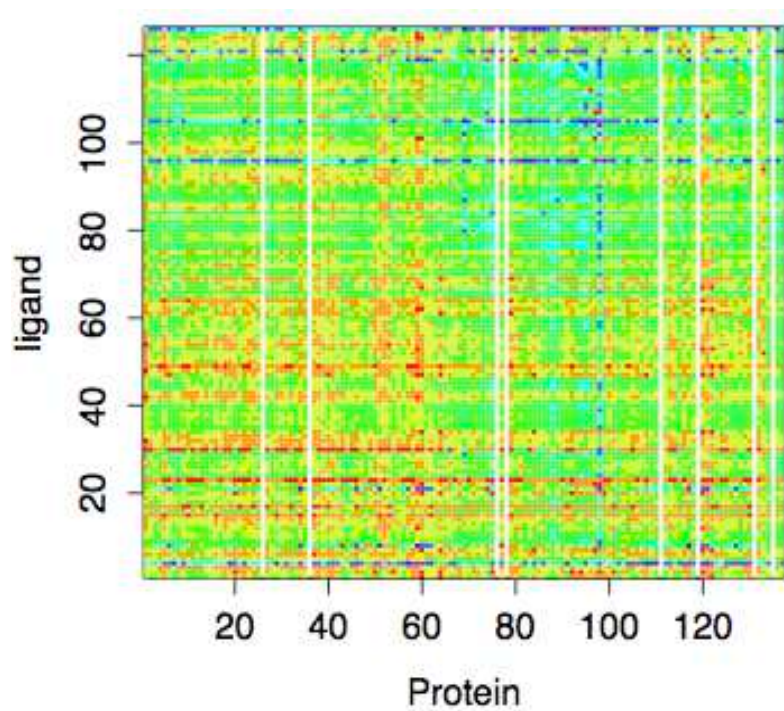


Figure 6.4. Heat map plot of SILE values representing the binding affinity of all the proteins against the various ligands.

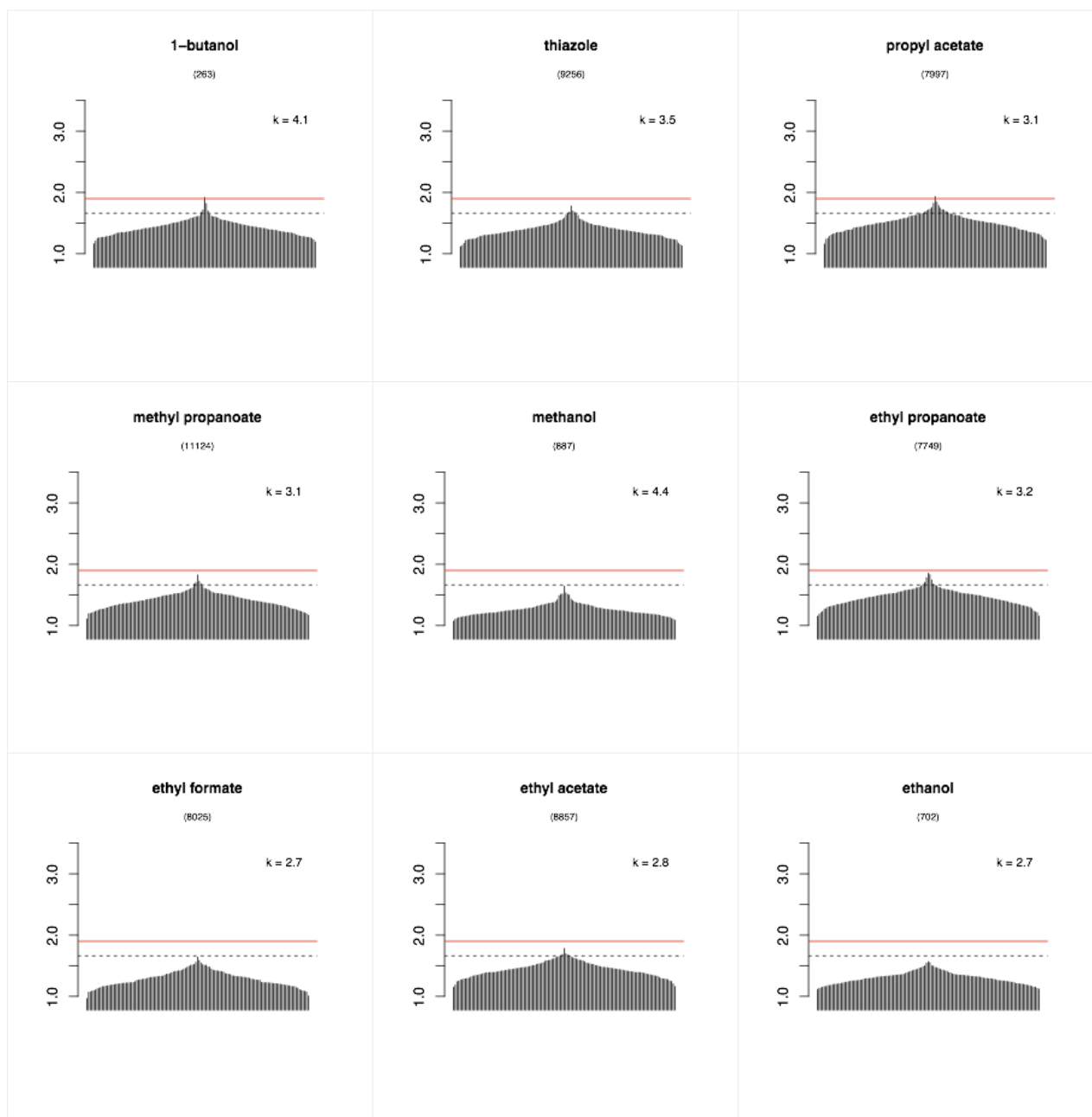


Figure 6.5. Odorant tuning curves of ligands which indicate low binding efficiency to OBPs. The SILE value are defined on the Y-axis and the OBPs on the X-axis. The highest SILE value for the given ligand for any OBP is plotted in the middle and are expanded on both the sides in the order of decreasing SILE values. Hence the order of receptor is not the same along the x-axis. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values (continued on next page).

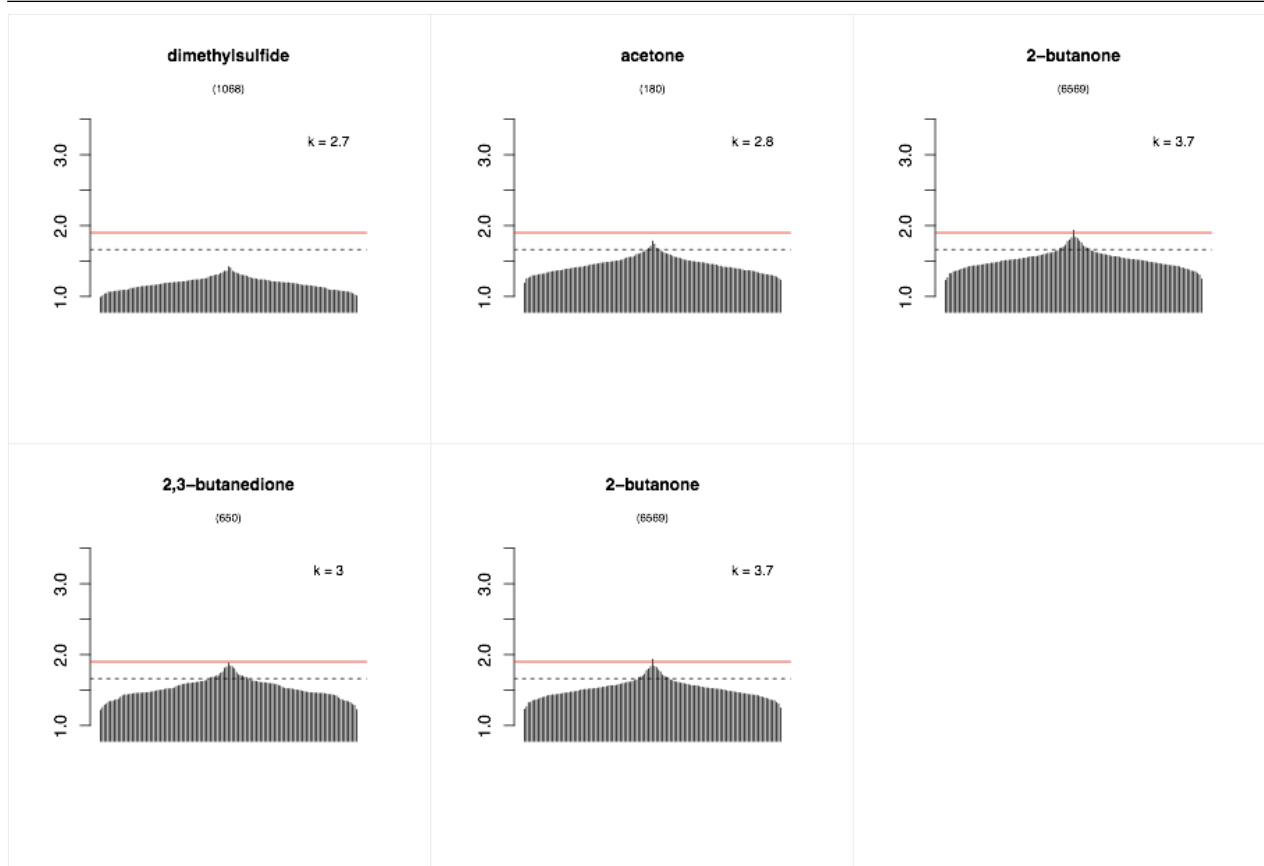


Figure 6.5 (contd). Odorant tuning curves of ligands which indicate low binding efficiency to OBPs. The SILE value are defined on the Y-axis and the OBPs on the X-axis. The highest SILE value for the given ligand for any OBP is plotted in the middle and are expanded on both the sides in the order of decreasing SILE values. Hence the order of receptor is not the same along the x-axis. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.

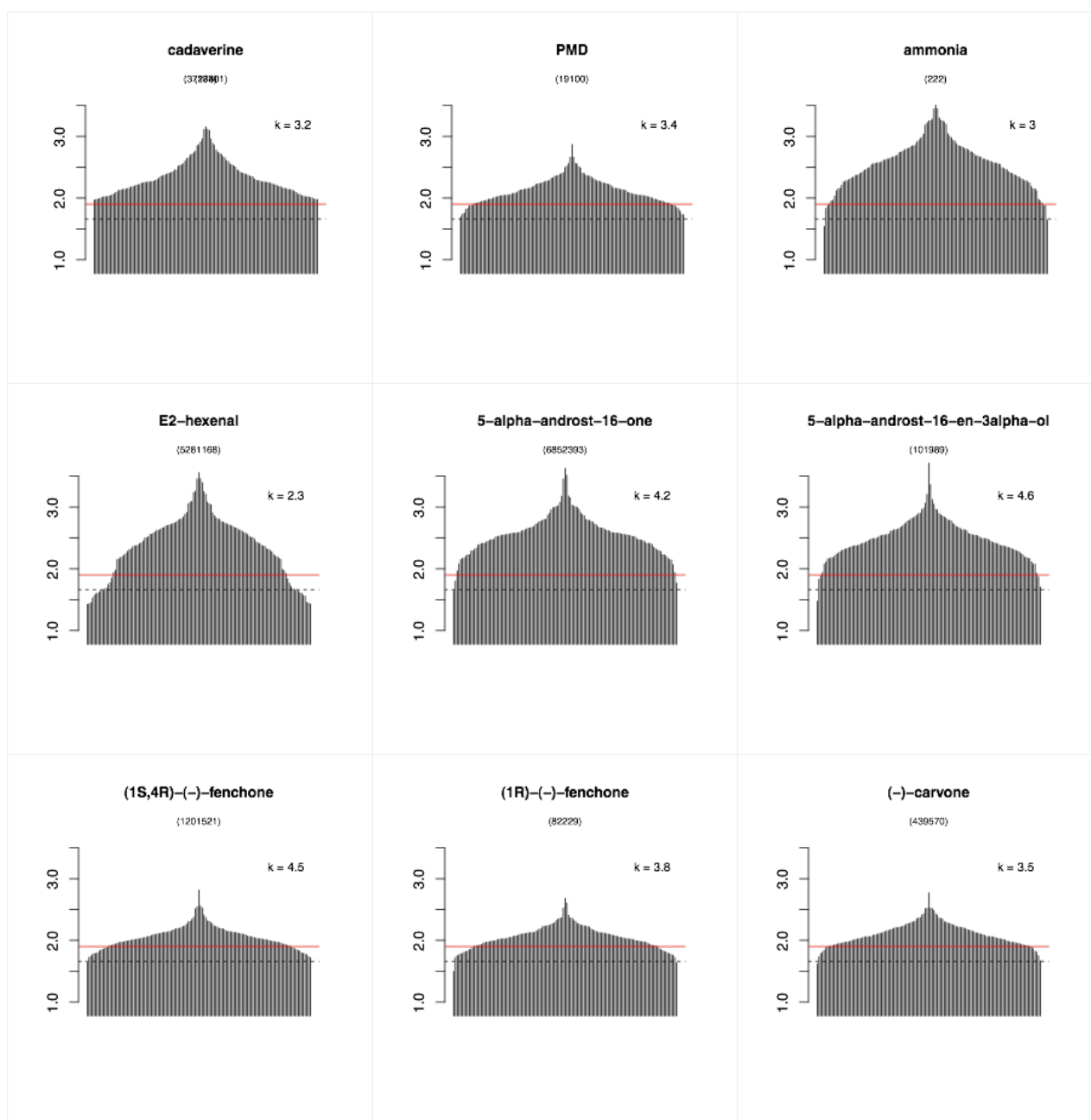


Figure 6.6. Ligands which indicate a broad spectrum of binding to OBPs. The SILE value are defined on the Y-axis and the receptors on the X-axis. The highest SILE value for the given ligand for any receptor is plotted in the middle and are expanded on both the sides in the order of decreasing SILE values. Hence the order of receptor is not the same along the X-axis. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values (continued on next page).

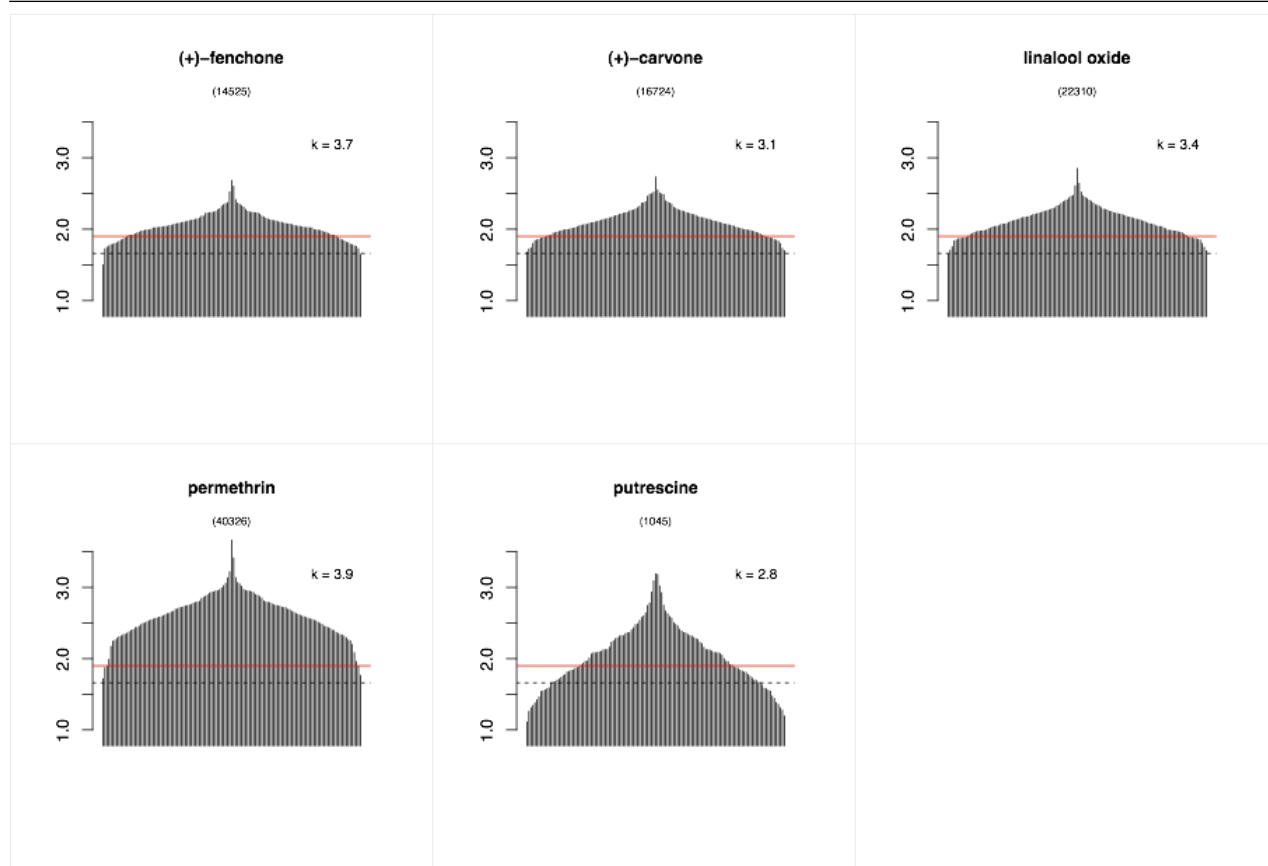


Figure 6.6 (contd). Ligands which indicate a broad spectrum of binding to OBPs. The SILE value are defined on the Y-axis and the receptors on the X-axis. The highest SILE value for the given ligand for any receptor is plotted in the middle and are expanded on both the sides in the order of decreasing SILE values. Hence the order of receptor is not the same along the X-axis. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.

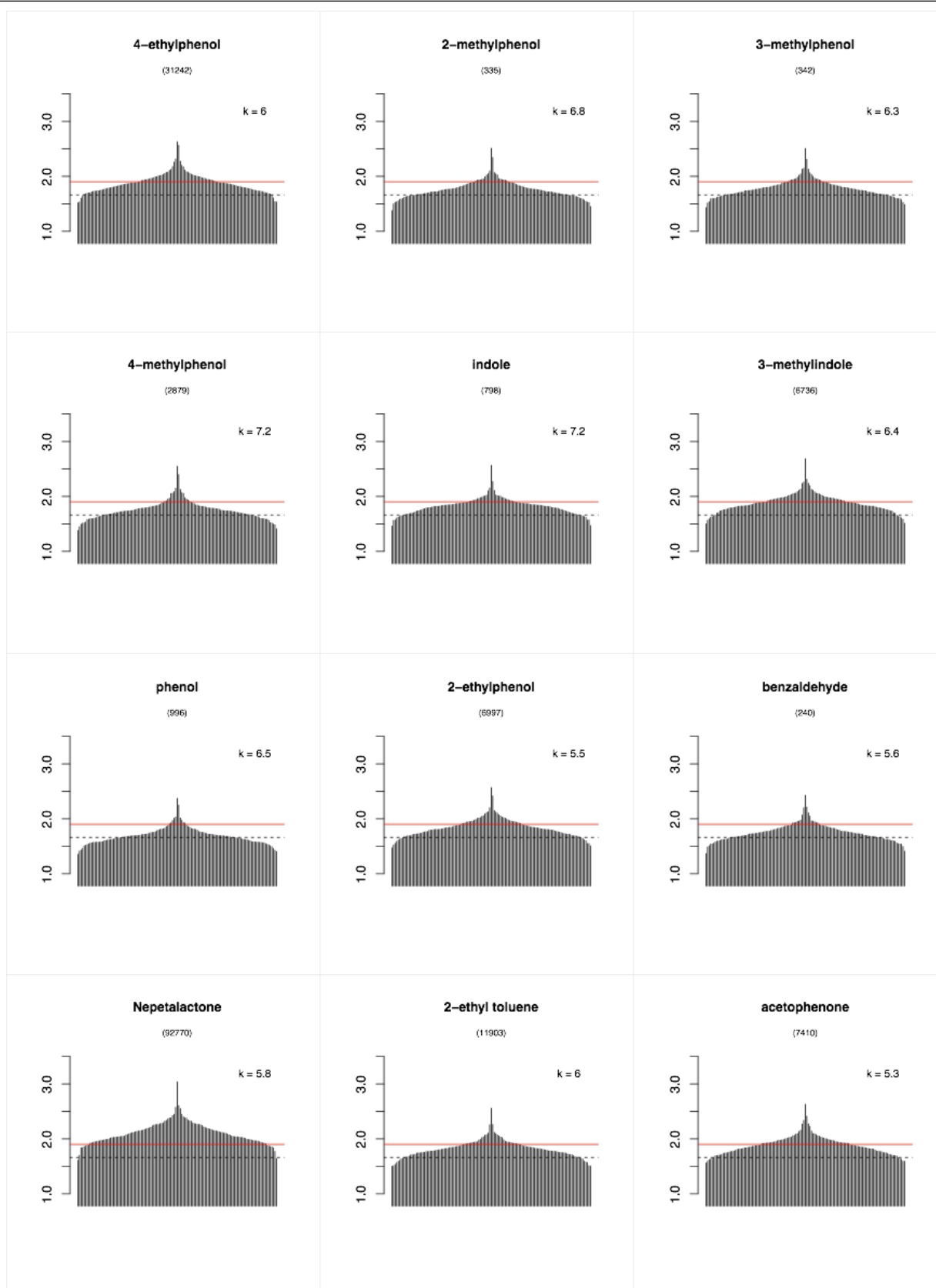


Figure 6.7. Tuning curves of the ligands which show specific binding to the OBPs. The SILE values are defined on the Y-axis and the receptors on the X-axis. The highest SILE value for the given ligand for any receptor is plotted in the middle and is expanded on both sides in the order of decreasing SILE values. Hence the order of receptor is not the same along the X-axis. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.

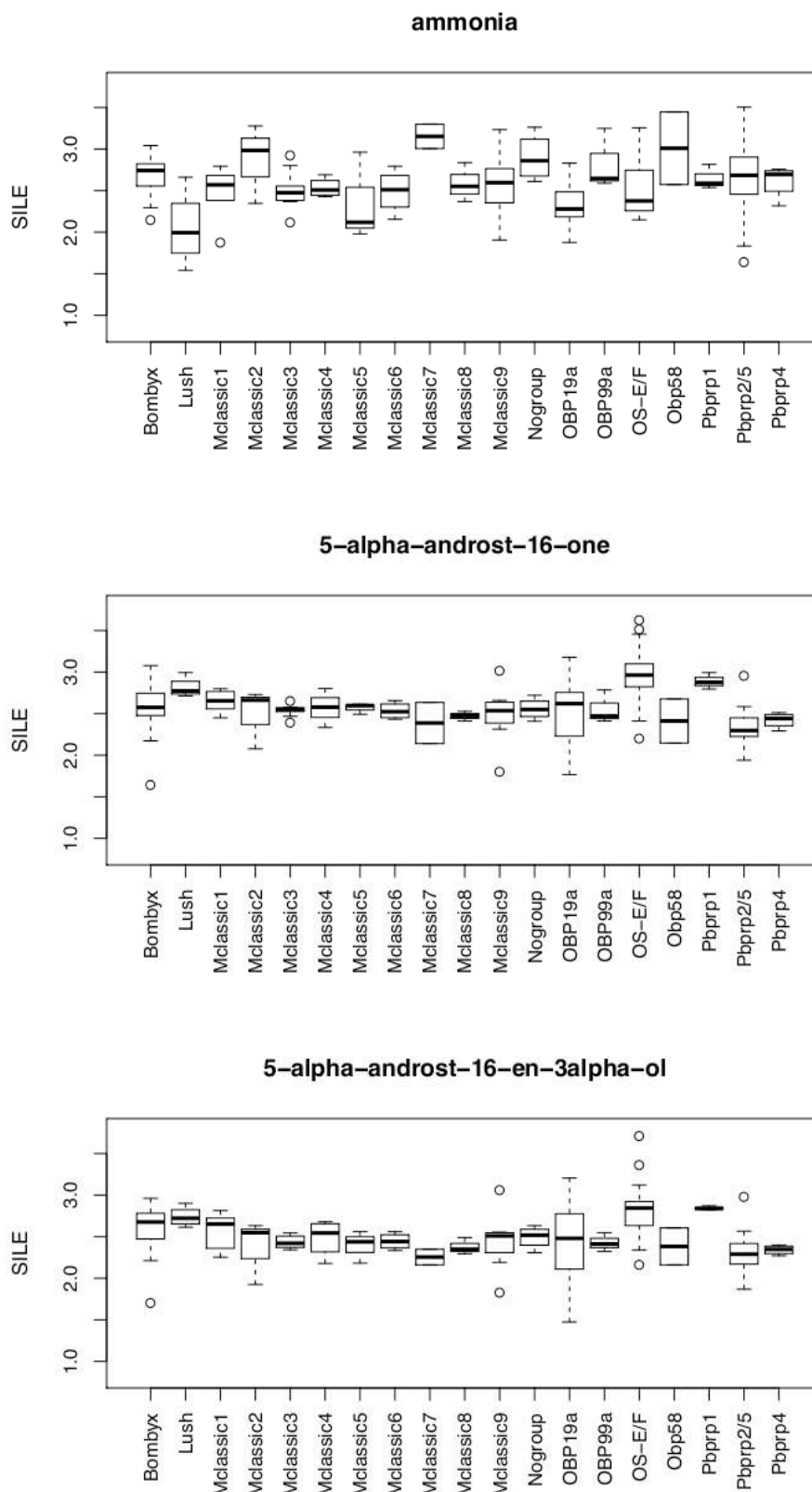


Figure 6.8. Box plot of the SILE values for broad spectrum ligands which show high binding affinity to OBPs in all the clusters. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.

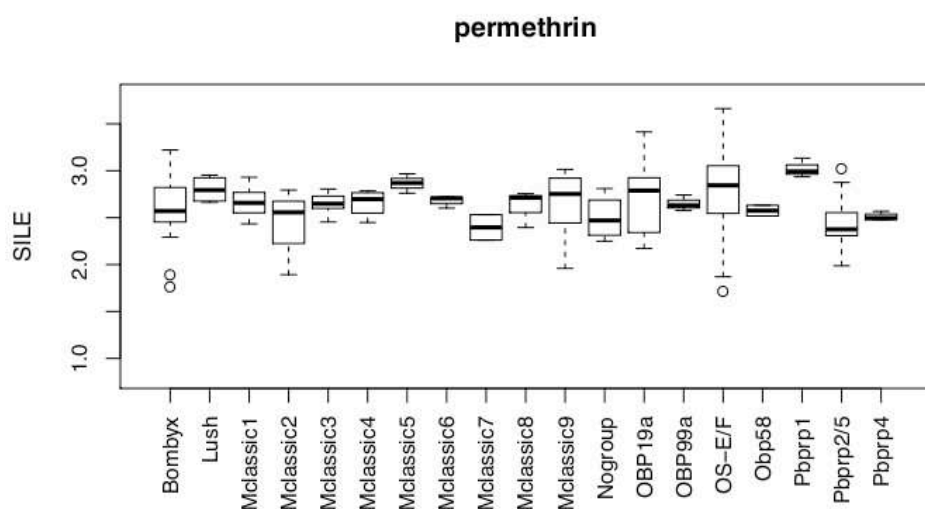


Figure 6.9. Box plot of the SILE values for the repellent permethrin which shows the highest binding affinity to all the clusters. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a give ligand across the clusters.

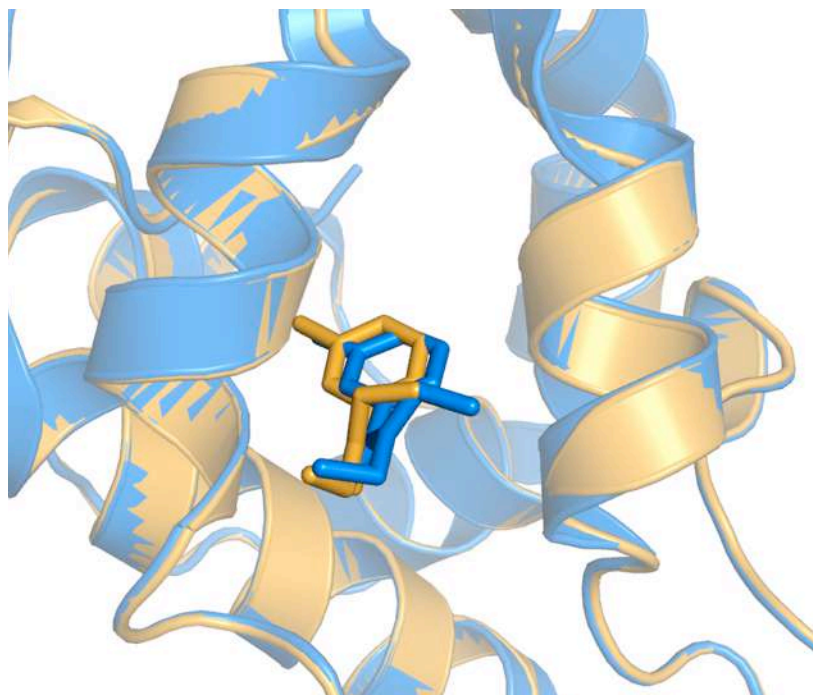


Figure 6.10. Structural superposition on DEET bound to AgamOBP1 in the crystal structure and the predicted complex using AUTODOCK. The ligand in the crystal structure is represented in blue and in the docked complex is represented in gold.

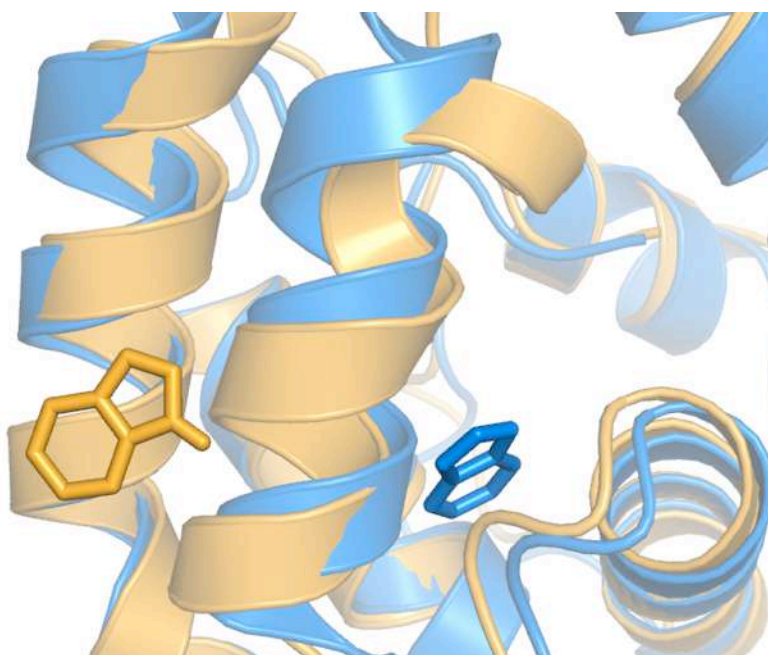


Figure 6.11. Structural superposition on indole bound to AgamOBP4 in the crystal structure and the predicted complex using AUTODOCK. The ligand in the crystal structure is represented in blue and in the docked complex is represented in orange.

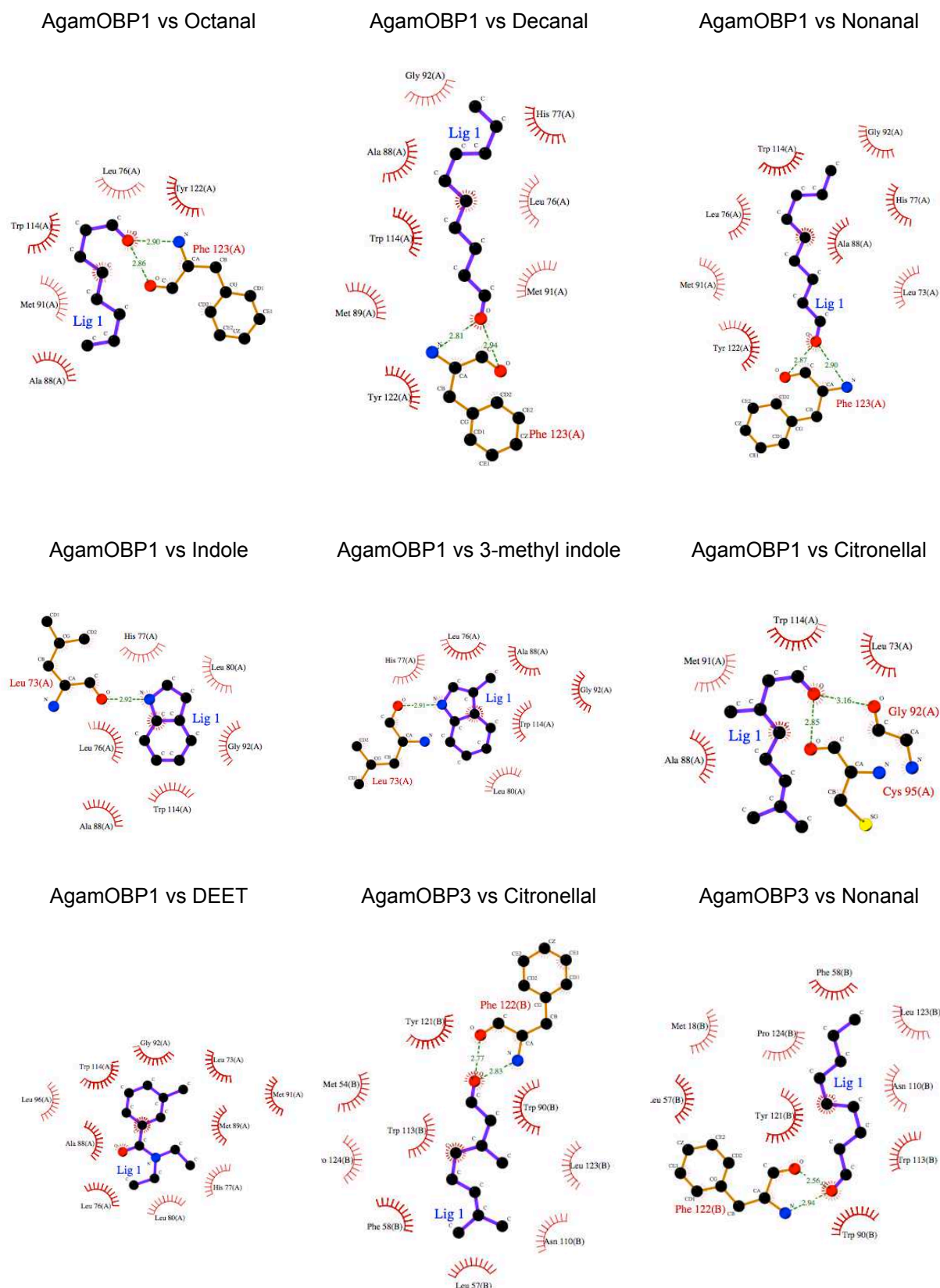
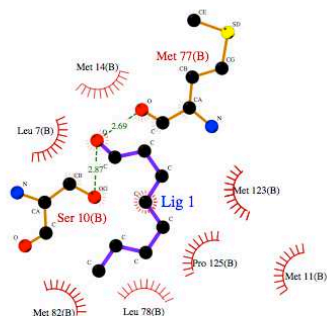
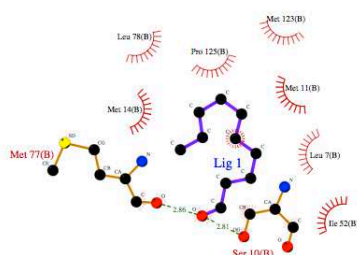


Figure 6.12. Analysis of the AUTODOCK results for a set of ligands that have been shown experimentally to bind AaegOBP1, AgamOBP1, AgamOBP4, AgamOBP19 and CquiOBP1. Shown are the LIGPLOT profiles representing the predicted interaction between the binding site residues and the corresponding ligand (continued on next page).

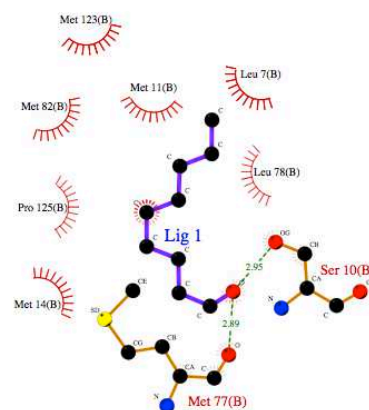
AgamOBP4 vs Octanal



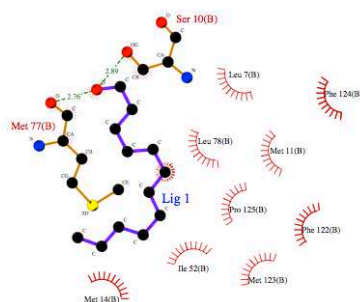
AgamOBP4 vs Decanal



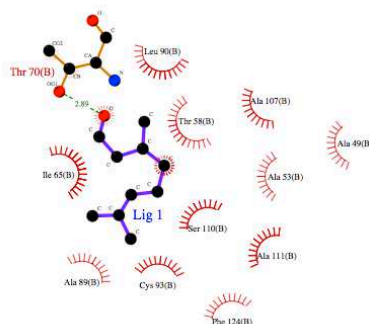
AgamOBP4 vs Nonanal



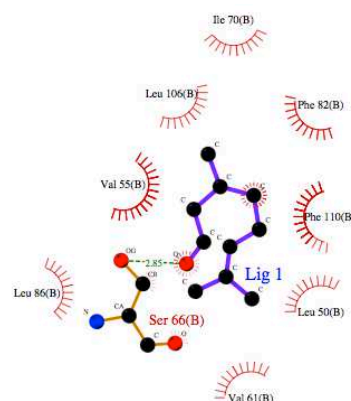
AgamOBP4 vs Dodecanal



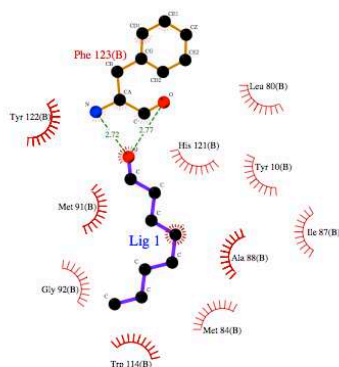
AgamOBP4 vs Citronellal



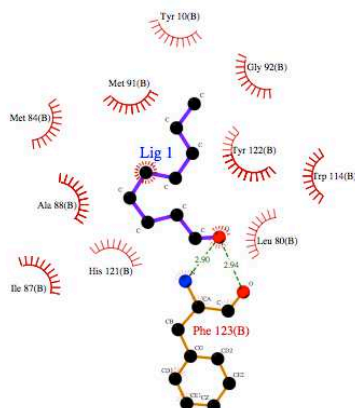
AgamOBP19 vs Citronellal



AaegOBP1 vs Octanal



AaegOBP1 vs Nonanal



AaegOBP1 vs Decanal

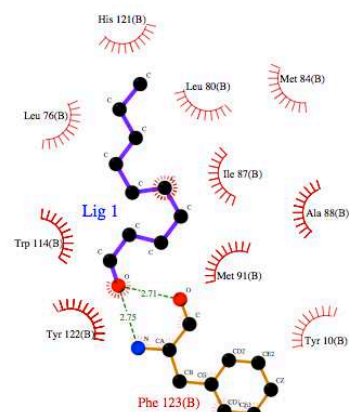
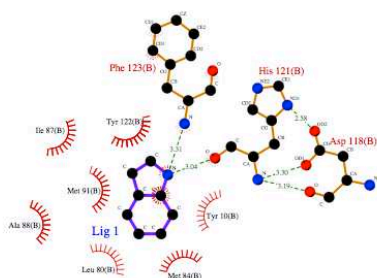
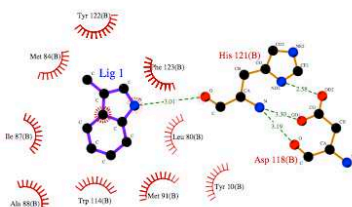


Figure 6.12 (contd). Analysis of the AUTODOCK results for a set of ligands that have been shown experimentally to bind AaegOBP1, AgamOBP1, AgamOBP4, AgamOBP19 and CquiOBP1. Shown are the LIGPLOT profiles representing the predicted interaction between the binding site residues and the corresponding ligand.

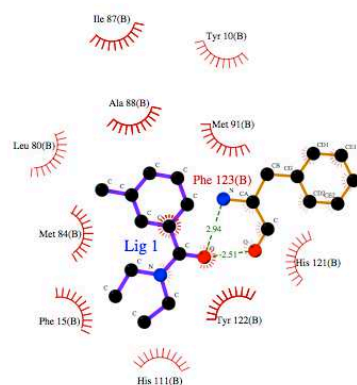
AaegOBP1 vs Indole



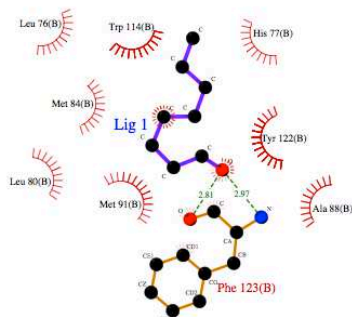
AaegOBP1 vs 3-methyl indole



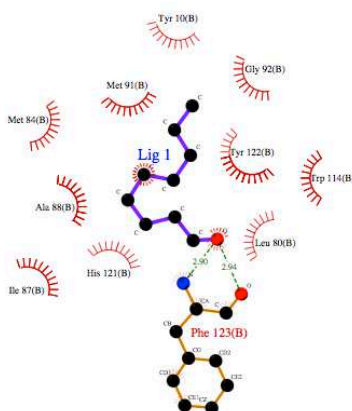
AaegOBP1 vs DEET



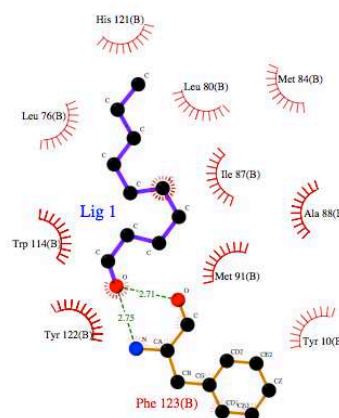
CquiOBP1 vs Octanal



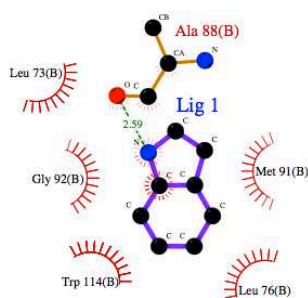
CquiOBP1 vs Nonanal



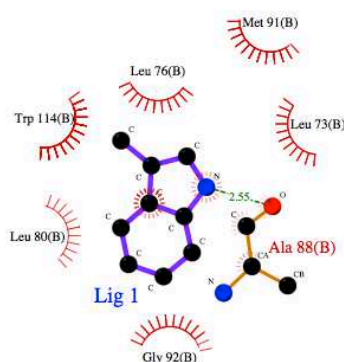
CquiOBP1 vs Decanal



CquiOBP1 vs Indole



CquiOBP1 vs 3-methyl indole



CquiOBP1 vs DEET

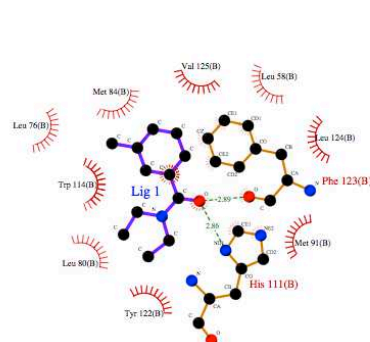


Figure 6.12 (contd). Analysis of the AUTODOCK results for a set of ligands that have been shown experimentally to bind AaegOBP1, AgamOBP1, AgamOBP4, AgamOBP19 and CquiOBP1. Shown are the LIGPLOT profiles representing the predicted interaction between the binding site residues and the corresponding ligand.

Table 6.1. Characteristics of the odorant molecules used in this study and overview of the results of the docking experiments performed using AutoDock (see materials and methods). For each odorant, shown are the predicted mean SILE values for the ligand across all OBPs, the top cluster and the mean SILE value for that cluster, the top5 odorant binding proteins and their SILE values.

Pubchem ID ¹	Compound ²	Formula ³	Type	mw ⁴	H ⁵ don.	H ⁶ acc.	Rotat bonds ⁷	dopp ⁸	Tau ⁹	Chrg ¹⁰	HA ¹¹	General mean SILE ¹²	Top cluster ¹³	avg of top cluster ¹⁴	Top #1	SILE	Top #2	SILE	Top #3	SILE	Top #4	SILE	Top #5	SILE	Hum. eman. ¹⁵	Ovip. ¹⁶	Bhv. ¹⁷	activity ¹⁸	PMID/DOI ¹⁹
439570	(-)-carvone	C10H14O	terpene	150	0	1	1	2.4	5	0	11	2.10±0.19	OS-E/F	2.34±0.28	aaegobp80 (Nogroup)	2.52	agamobp6 (OBP19a)	2.52	cquiobp2 (OS-E/F)	2.51	agamobp19 (OBP19a)	2.51	cquiobp13 (OBP19a)	2.49				repellent	8406635
16724	(+)-carvone	C10H14O	terpene	150	0	1	1	2.4	5	0	11	2.10±0.20	OS-E/F	2.35±0.28	cquiobp13 (OBP19a)	2.55	cquiobp4 (OS-E/F)	2.52	agamobp6 (OBP19a)	2.51	cquiobp2 (OS-E/F)	2.50	cquiobp5 (OS-E/F)	2.49				repellent	8406635
14525	(+)-fenchone	C10H16O	terpene	152	0	1	0	2.3	1	0	11	2.05±0.19	OS-E/F	2.29±0.27	cquiobp13 (OBP19a)	2.60	agamobp6 (OBP19a)	2.52	cquiobp4 (OS-E/F)	2.41	aaegobp38 (OS-E/F)	2.37	aaegobp78 (Bombyx)	2.36				repellent	12428949
7793	(±)-beta-Citronellol	C10H20O	terpene	156	1	1	5	3.2	1	0	11	1.86±0.23	OS-E/F	2.21±0.27	agamobp6 (OBP19a)	2.44	cquiobp2 (OS-E/F)	2.42	cquiobp58 (OS-E/F)	2.36	aaegobp38 (OS-E/F)	2.30	cquiobp4 (OS-E/F)	2.25				repellent	8406635
82229	(1R)-(-)-fenchone	C10H16O	terpene	152	0	1	0	2.3	1	0	11	2.05±0.19	OS-E/F	2.29±0.28	cquiobp13 (OBP19a)	2.60	agamobp6 (OBP19a)	2.52	cquiobp4 (OS-E/F)	2.41	aaegobp38 (OS-E/F)	2.37	aaegobp78 (Bombyx)	2.36				repellent	12428949
1201521	(1S,4R)-(-)-fenchone	C10H16O	terpene	152	0	1	0	2.3	1	0	11	2.06±0.19	OS-E/F	2.31±0.22	cquiobp13 (OBP19a)	2.56	agamobp6 (OBP19a)	2.54	cquiobp58 (OS-E/F)	2.52	aaegobp38 (OS-E/F)	2.51	cquiobp4 (OS-E/F)	2.41				repellent	12428949
263	1-butanol	C4H10O	alcohol	74.1	1	1	2	0.9	1	0	5	1.43±0.13	OS-E/F	1.58±0.15	cquiobp58 (OS-E/F)	1.82	cquiobp2 (OS-E/F)	1.71	aaegobp56 (OBP19a)	1.70	cquiobp13 (OBP19a)	1.67	cquiobp69 (Bombyx)	1.64	yes			attractant	20017925
8192	1-chlorododecane	C12H25Cl	alkane	205	0	0	10	6.9	1	0	13	1.79±0.23	OS-E/F	2.12±0.28	cquiobp2 (OS-E/F)	2.39	cquiobp58 (OS-E/F)	2.35	aaegobp38 (OS-E/F)	2.24	cquiobp13 (OBP19a)	2.24	cquiobp4 (OS-E/F)	2.21	yes			attractant	10701259
8193	1-dodecanol	C12H26O	alcohol	186	1	1	10	5.1	1	0	13	1.74±0.24	OS-E/F	2.11±0.28	cquiobp2 (OS-E/F)	2.39	aaegobp38 (OS-E/F)	2.26	agamobp2 (OS-E/F)	2.22	cquiobp13 (OBP19a)	2.22	cquiobp58 (OS-E/F)	2.21	yes			attractant	10.1023/A:1005475422978
21057	1-hepten-3-ol	C7H14O	alcohol	114	1	1	4	2	1	0	8	1.65±0.18	OS-E/F	1.89±0.36	cquiobp58 (OS-E/F)	2.20	cquiobp13 (OBP19a)	2.13	aaegobp56 (OBP19a)	2.03	cquiobp2 (OS-E/F)	2.01	agamobp6 (OBP19a)	2.00	yes			attractant	10701259
8103	1-hexanol	C6H14O	alcohol	102	1	1	4	2	1	0	7	1.53±0.18	OS-E/F	1.76±0.24	cquiobp58 (OS-E/F)	2.02	cquiobp13 (OBP19a)	1.96	aaegobp56 (OBP19a)	1.91	cquiobp2 (OS-E/F)	1.88	cquiobp5 (OS-E/F)	1.88	yes			repellent	19710651
20928	1-hexen-3-ol	C6H12O	alcohol	100	1	1	3	1.5	1	0	7	1.59±0.16	OS-E/F	1.81±0.18	cquiobp58 (OS-E/F)	2.07	cquiobp2 (OS-E/F)	1.97	cquiobp13 (OBP19a)	1.96	aaegobp56 (OBP19a)	1.92	cquiobp5 (OS-E/F)	1.88	yes			attractant	10701259
18827	1-octen-3-ol	C8H16O	alcohol	128	1	1	5	2.6	1	0	9	1.69±0.20	OS-E/F	1.98±0.26	cquiobp58 (OS-E/F)	2.23	cquiobp13 (OBP19a)	2.23	cquiobp2 (OS-E/F)	2.13	cquiobp5 (OS-E/F)	2.11	cquiobp4 (OS-E/F)	2.09	yes			attractant	8887339
6276	1-pentanol	C5H12O	alcohol	88.2	1	1	3	1.6	1	0	6	1.48±0.15	OS-E/F	1.66±0.18	cquiobp58 (OS-E/F)	1.92	cquiobp13 (OBP19a)	1.83	aaegobp56 (OBP19a)	1.81	cquiobp2 (OS-E/F)	1.76	cquiobp69 (Bombyx)	1.75	yes			attractant	19710651
14286	2-acetylpyridine	C7H7NO	heterocyclic	121	0	2	1	0.9	2	0	9	1.79±0.16	OS-E/F	1.94±0.20	cquiobp13 (OBP19a)	2.21	aaegobp56 (OBP19a)	2.17	cquiobp60 (Bombyx)	2.16	cquiobp58 (OS-E/F)	2.15	cquiobp2 (OS-E/F)	2.08				attractant	20017925
520108	2-acetylthiazole	C5H5NOS	heterocyclic	127	0	1	1	1	2	0	8	1.70±0.15	OS-E/F	1.82±0.19	cquiobp60 (Bombyx)	2.08	aaegobp56 (OBP19a)	2.06	cquiobp13 (OBP19a)	2.05	cquiobp58 (OS-E/F)	2.04	cquiobp2 (OS-E/F)	1.97					16938890
6920	2-acetylthiophene	C6H6OS	heterocyclic	126	0	1	1	1.2	2	0	8	1.79±0.16	OS-E/F	1.97±0.18	cquiobp13 (OBP19a)	2.24	cquiobp58 (OS-E/F)	2.22	aaegobp56 (OBP19a)	2.20	cquiobp2 (OS-E/F)	2.13	agamobp6 (OBP19a)	2.07					16938890
6569	2-butanone	C4H8O	ketone	72.1	0	1	1	0.3	3	0	5	1.52±0.13	OS-E/F	1.62±0.20	cquiobp27 (Pbprp2/5)	1.83	cquiobp58 (OS-E/F)	1.83	cquiobp2 (OS-E/F)	1.82	cquiobp54 (Pbprp2/5)	1.80	aaegobp13 (Mclassic4)	1.78	yes			attractant	10701259
61809	2-ethoxythiazole	C5H7NOS	heterocyclic	129	0	1	2	1.6	1	0	8	1.57±0.14	Pbprp1	1.70±0.17	cquiobp2 (OS-E/F)	1.97	cquiobp58 (OS-E/F)	1.87	cquiobp13 (OBP19a)	1.86	cquiobp5 (OS-E/F)	1.83	aaegobp2 (Pbprp1)	1.82					20160092

¹Pubchem ID, ²Compound name, ³chemical formula, ⁴molecular weight, ⁵n° of hydrogen donor atoms, ⁶n° of hydrogen acceptor atoms, ⁷n° of rotatable bonds, ⁸partition coefficient, ⁹n° of tautomers, ¹⁰charge, ¹¹n° of heavy atoms, ¹²⁻¹⁴results of docking experiments, ¹²mean ligand efficiency for the ligand (SILE), ¹³cluster having best SILE value, ¹⁴average SILE value of best cluster, ¹⁵Presence in human emanation, ¹⁶Affect oviposition behavior, ¹⁷Affect other behaviors of mosquitoes, ¹⁸known activity of the compound, ¹⁹Pubmed ID or DOI number.

Pubchem ID ¹	Compound ²	Formula ³	Type	mw ⁴	H ⁵ don.	H ⁶ acc.	Rotat bonds ⁷	logp ⁸	Tau ⁹	Chrg ¹⁰	HA ¹¹	General mean SILE ¹²	Top cluster ¹³	avg of top cluster ¹⁴	Top #1	SILE	Top #2	SILE	Top #3	SILE	Top #4	SILE	Top #5	SILE	Hum. eman. ¹⁵	Ovip. ¹⁶	Bhv. ¹⁷	activity ¹⁸	PMID/DOI ¹⁹
11903	2-ethyl toluene	C9H12	aromatic	120	0	0	1	3.5	1	0	9	1.82±0.16	OS-E/F	2.01±0.31	cquiobp58 (OS-E/F)	2.26	cquiobp13 (OBP19a)	2.26	agamobp6 (OBP19a)	2.12	cquiobp2 (OS-E/F)	2.10	agamobp7 (Pbprp1)	2.09					
7720	2-ethyl-1-hexanol	C8H18O	alcohol	130	1	1	5	3.1	1	0	9	1.68±0.20	OS-E/F	1.99±0.24	cquiobp58 (OS-E/F)	2.20	cquiobp2 (OS-E/F)	2.14	agamobp6 (OBP19a)	2.14	aaegobp56 (OBP19a)	2.13	cquiobp13 (OBP19a)	2.11					
6997	2-ethylphenol	C8H10O	aromatic	122	1	1	1	2.5	3	0	9	1.83±0.17	OS-E/F	2.05±0.20	cquiobp58 (OS-E/F)	2.42	cquiobp13 (OBP19a)	2.20	agamobp6 (OBP19a)	2.15	cquiobp61 (Bombyx)	2.13	cquiobp2 (OS-E/F)	2.12					
8051	2-heptanone	C7H14O	ketone	114	0	1	4	2	3	0	8	1.68±0.17	OS-E/F	1.90±0.27	cquiobp58 (OS-E/F)	2.15	aaegobp56 (OBP19a)	2.13	cquiobp13 (OBP19a)	2.08	cquiobp2 (OS-E/F)	2.04	cquiobp5 (OS-E/F)	1.98	yes				10431355
62725	2-iso-butyl-thiazole	C7H11NS	heterocyclic	141	0	0	2	2.6	2	0	9	1.79±0.17	OS-E/F	2.00±0.21	cquiobp13 (OBP19a)	2.20	cquiobp5 (OS-E/F)	2.20	cquiobp2 (OS-E/F)	2.18	cquiobp58 (OS-E/F)	2.16	aaegobp38 (OS-E/F)	2.15					
335	2-methylphenol	C7H8O	aromatic	108	1	1	0	2	3	0	8	1.76±0.16	OS-E/F	1.97±0.21	cquiobp58 (OS-E/F)	2.34	agamobp6 (OBP19a)	2.10	cquiobp13 (OBP19a)	2.07	cquiobp61 (Bombyx)	2.05	aaegobp38 (OS-E/F)	2.04					
13187	2-nonanone	C9H18O	ketone	142	0	1	6	3.1	3	0	10	1.76±0.21	OS-E/F	2.06±0.32	cquiobp2 (OS-E/F)	2.26	cquiobp58 (OS-E/F)	2.23	aaegobp56 (OBP19a)	2.17	aaegobp38 (OBP19a)	2.16	cquiobp13 (OBP19a)	2.15	yes	yes		attractant	19058627
58	2-oxobutanoic acid	C4H6O3	acid	102	1	3	2	0.1	2	0	7	1.63±0.26	mclassic2	1.86±0.42	agamobp19 (OBP19a)	2.40	cquiobp28 (Pbprp2/5)	2.24	agamobp9 (OBP99a)	2.22	cquiobp61 (Bombyx)	2.19	cquiobp29 (Pbprp2/5)	2.17			yes	attractant	12109705
6419709	2-oxohexanoic acid	C6H9O3-	acid	129	0	3	3	1.7	2	-1	9	1.60±0.24	OBP99a	1.87±0.39	agamobp9 (OBP99a)	2.32	aaegobp77 (Mclassic2)	2.18	cquiobp4 (OS-E/F)	2.12	cquiobp61 (Bombyx)	2.10	cquiobp37 (Pbprp2/5)	2.09			yes	attractant	12109705
74563	2-oxopentanoic acid	C5H8O3	acid	116	1	3	3	0.5	2	0	8	1.64±0.26	mclassic2	1.91±0.36	agamobp19 (OBP19a)	2.42	cquiobp28 (Pbprp2/5)	2.27	cquiobp29 (Pbprp2/5)	2.20	cquiobp61 (Bombyx)	2.19	agamobp9 (OBP99a)	2.18			yes	attractant	12109705
1060	2-oxopropanoic acid	C3H4O3	acid	88.1	1	3	1	-0.3	2	0	6	1.63±0.27	mclassic8	1.85±0.51	agamobp19 (OBP19a)	2.44	agamobp9 (OBP99a)	2.29	cquiobp29 (Pbprp2/5)	2.24	aaegobp18 (Mclassic9)	2.21	cquiobp61 (Bombyx)	2.21	yes			attractant	12109705
31236	2-phenoxy ethanol	C8H10O2	aromatic	138	1	2	3	1.2	1	0	10	1.87±0.18	OS-E/F	2.08±0.31	cquiobp58 (OS-E/F)	2.30	cquiobp2 (OS-E/F)	2.29	cquiobp4 (OS-E/F)	2.29	agamobp65 (Nogroup)	2.24	cquiobp5 (OS-E/F)	2.23					
12570	2-propylphenol	C9H12O	aromatic	136	1	1	2	2.9	3	0	10	1.89±0.18	OS-E/F	2.14±0.25	cquiobp58 (OS-E/F)	2.47	cquiobp13 (OBP19a)	2.30	cquiobp2 (OS-E/F)	2.27	cquiobp5 (OS-E/F)	2.27	agamobp6 (OBP19a)	2.23					
650	2,3-butanedione	C4H6O2	ketone	86.1	0	2	1	-1.3	3	0	6	1.52±0.13	OS-E/F	1.60±0.19	cquiobp29 (Pbprp2/5)	1.84	cquiobp2 (OS-E/F)	1.82	cquiobp58 (OS-E/F)	1.82	aaegobp18 (Mclassic9)	1.81	cquiobp70 (Bombyx)	1.80	yes			attractant	20017925
61653	2,4,5-trimethyl thiazole	C6H9NS	heterocyclic	127	0	0	0	2.2	1	0	8	1.77±0.14	Pbprp1	1.91±0.22	cquiobp58 (OS-E/F)	2.18	cquiobp13 (OBP19a)	2.12	aaegobp80 (Nogroup)	2.06	cquiobp67 (Bombyx)	2.05	agamobp7 (Pbprp1)	2.03					
31260	3-methyl-1-butanol	C5H12O	alcohol	88.2	1	1	2	1.2	1	0	6	1.54±0.15	OS-E/F	1.72±0.18	cquiobp58 (OS-E/F)	1.98	cquiobp13 (OBP19a)	1.92	aaegobp56 (OBP19a)	1.87	agamobp6 (OBP19a)	1.84	cquiobp2 (OS-E/F)	1.82	yes	yes	yes	repellent	16963500
89487	3-methyl-2-cyclohexen-1-ol	C7H12O	alcohol	112	1	1	0	0.9	1	0	8	1.87±0.18	OS-E/F	2.11±0.21	cquiobp58 (OS-E/F)	2.40	cquiobp13 (OBP19a)	2.38	aaegobp56 (OBP19a)	2.29	aaegobp38 (OS-E/F)	2.25	agamobp6 (OBP19a)	2.22					
6443739	3-methyl-2-hexenoic acid	C7H12O2	acid	128	1	2	3	2.2	1	0	9	1.69±0.23	mclassic2	2.00±0.45	aaegobp77 (Mclassic2)	2.28	agamobp24 (Mclassic2)	2.24	cquiobp61 (Bombyx)	2.23	cquiobp28 (Pbprp2/5)	2.19	agamobp14 (Mclassic9)	2.14	yes			attractant	16690870
6736	3-methylindole	C9H9N	aromatic	131	1	0	0	2.6	1	0	10	1.88±0.17	OS-E/F	2.08±0.23	cquiobp58 (OS-E/F)	2.31	cquiobp13 (OBP19a)	2.26	agamobp6 (OBP19a)	2.24	cquiobp2 (OS-E/F)	2.23	aaegobp79 (Nogroup)	2.19	yes	yes		attractant	1583482
342	3-methylphenol	C7H8O	aromatic	108	1	1	0	2	3	0	8	1.78±0.16	OS-E/F	1.99±0.19	cquiobp58 (OS-E/F)	2.31	cquiobp13 (OBP19a)	2.14	agamobp6 (OBP19a)	2.13	aaegobp56 (OBP19a)	2.13	cquiobp2 (OS-E/F)	2.06					
246728	3-octanone	C8H16O	ketone	128	0	1	5	2.3	3	0	9	1.71±0.19	OS-E/F	1.96±0.26	cquiobp58 (OS-E/F)	2.24	cquiobp13 (OBP19a)	2.16	cquiobp2 (OS-E/F)	2.15	cquiobp5 (OS-E/F)	2.11	cquiobp4 (OS-E/F)	2.03					
31242	4-ethylphenol	C8H10O	aromatic	122	1	1	1	2.6	2	0	8	1.88±0.18	OS-E/F	2.12±0.23	cquiobp58 (OS-E/F)	2.57	cquiobp5 (OS-E/F)	2.31	cquiobp2 (OS-E/F)	2.27	cquiobp13 (OBP19a)	2.26	agamobp6 (OBP19a)	2.20		yes	yes	repellent	16963500

¹Pubchem ID, ²Compound name, ³chemical formula, ⁴molecular weight, ⁵n° of hydrogen donor atoms, ⁶n° of hydrogen acceptor atoms, ⁷n° of rotatable bonds, ⁸partition coefficient, ⁹n° of tautomers, ¹⁰charge, ¹¹n° of heavy atoms, ¹²⁻¹⁴results of docking experiments, ¹²mean ligand efficiency for the ligand (SILE), ¹³cluster having best SILE value, ¹⁴average SILE value of best cluster, ¹⁵Presence in human emanation, ¹⁶Affect oviposition behavior, ¹⁷Affect other behaviors of mosquitoes, ¹⁸known activity of the compound, ¹⁹Pubmed ID or DOI number.

Pubchem ID ¹	Compound ²	Formula ³	Type	mw ⁴	H ⁵ don.	H ⁶ acc.	Rotat bonds ⁷	dlogp ⁸	Tau ⁹	Chrg ¹⁰	HA ¹¹	General mean SILE ¹²	Top cluster ¹³	avg of top cluster ¹⁴	Top #1	SILE	Top #2	SILE	Top #3	SILE	Top #4	SILE	Top #5	SILE	Hum. eman. ¹⁵	Ovip. ¹⁶	Bhv. ¹⁷	activity ¹⁸	PMID/DOI ¹⁹
11524	4-methylcyclohexanol	C7H14O	alcohol	114	1	1	0	1.8	1	0	8	1.89±0.18	OS-E/F	2.16±0.24	cquiobp58 (OS-E/F)	2.57	agamobp6 (OBP19a)	2.32	cquiobp2 (OS-E/F)	2.30	cquiobp13 (OBP19a)	2.27	cquiobp5 (OS-E/F)	2.22		yes		attractant	7143381
2879	4-methylphenol	C7H8O	aromatic	108	1	1	0	1.9	2	0	8	1.75±0.17	OS-E/F	1.98±0.22	cquiobp58 (OS-E/F)	2.40	agamobp6 (OBP19a)	2.15	cquiobp2 (OS-E/F)	2.13	cquiobp13 (OBP19a)	2.09	aaegobp17 (Mclassic7)	2.07	yes	yes		attractant	7143381
12748	4-methylthiazole	C4H5NS	heterocyclic	99.2	0	0	0	1	1	0	6	1.54±0.13	OS-E/F	1.66±0.19	cquiobp58 (OS-E/F)	1.85	cquiobp2 (OS-E/F)	1.85	cquiobp70 (Bombyx)	1.82	aaegobp38 (OS-E/F)	1.80	cquiobp34 (Pbprp2/5)	1.79					
62510	4,5-dimethyl thiazole	C5H7NS	heterocyclic	113	0	0	0	1.8	1	0	7	1.64±0.14	OS-E/F	1.78±0.20	cquiobp58 (OS-E/F)	2.06	cquiobp70 (Bombyx)	2.00	cquiobp13 (OBP19a)	1.93	aaegobp80 (Nogroup)	1.88	aaegobp38 (OS-E/F)	1.88					
101989	5-alpha-androst-16-en-3alpha-ol	C19H30O	other	274	1	1	0	5.3	1	0	20	2.50±0.32	Pbprp1	2.84±0.49	cquiobp4 (OS-E/F)	3.36	agamobp19 (OBP19a)	3.21	aaegobp38 (OS-E/F)	3.12	cquiobp46 (Mclassic9)	3.06	agamobp6 (OBP19a)	3.05	yes			attractant	4416149
6852393	5-alpha-androst-16-one	C19H28O	other	272	0	1	0	4.9	3	0	20	2.57±0.32	OS-E/F	2.96±0.41	cquiobp4 (OS-E/F)	3.52	aaegobp38 (OS-E/F)	3.46	agamobp6 (OBP19a)	3.18	agamobp19 (OS-E/F)	3.18	agamobp19 (OBP19a)	3.15	yes			attractant	10.1007/ BF0098377 7
9862	6-methylhept-5-en-2-one	C8H14O	ketone	126	0	1	3	1.9	3	0	9	1.81±0.18	OS-E/F	2.06±0.24	cquiobp13 (OBP19a)	2.24	cquiobp58 (OS-E/F)	2.22	cquiobp2 (OS-E/F)	2.22	aaegobp38 (OS-E/F)	2.17	cquiobp5 (OS-E/F)	2.15	yes		yes	repellent	16963500
543921	7-octenoic acid	C9H16O2	acid	156	0	2	7	3.1	1	0	11	1.66±0.20	OS-E/F	1.92±0.28	cquiobp2 (OS-E/F)	2.15	cquiobp58 (OS-E/F)	2.14	cquiobp5 (OS-E/F)	2.06	aaegobp38 (OS-E/F)	2.05	agamobp27 (Mclassic1)	2.02	yes		yes	attractant	11583442
176	acetic acid	C2H4O2	acid	60.1	1	2	0	-0.2	1	0	4	1.55±0.27	mclassic8	1.78±0.62	agamobp9 (OBP99a)	2.36	agamobp19 (OBP19a)	2.25	cquiobp61 (Bombyx)	2.18	cquiobp28 (Pbprp2/5)	2.17	agamobp26 (Mclassic3)	2.03	yes		yes	repellent	10701259
180	acetone	C3H6O	ketone	58.1	0	1	0	-0.1	2	0	4	1.44±0.12	Pbprp1	1.62±0.17	agamobp7 (Pbprp1)	1.78	cquiobp27 (Pbprp2/5)	1.71	cquiobp2 (OS-E/F)	1.68	aaegobp56 (OBP19a)	1.67	cquiobp58 (OS-E/F)	1.67	yes		yes	attractant	10.1007/ BF0276560 6
7410	acetophenone	C8H8O	aromatic	120	0	1	1	1.6	2	0	9	1.88±0.17	OS-E/F	2.10±0.19	cquiobp13 (OBP19a)	2.41	cquiobp58 (OS-E/F)	2.34	cquiobp2 (OS-E/F)	2.27	aaegobp56 (OBP19a)	2.27	agamobp6 (OBP19a)	2.23					
222	ammonia	H3N	amine	17	1	1	0	-0.7	1	0	1	2.57±0.39	mclassic7	3.15±0.62	cquiobp31 (Pbprp2/5)	3.50	cquiobp54 (Pbprp2/5)	3.45	agamobp29 (OBP59a)	3.45	cquiobp53 (Mclassic7)	3.30	cquiobp21 (Mclassic2)	3.28	yes		yes	attractant	10.1046/ j.1365-3032. 2001.0022 7.x
12348	amyl acetate	C7H14O2	ester	130	0	2	5	1.9	1	0	9	1.61±0.18	OS-E/F	1.82±0.22	cquiobp13 (OBP19a)	1.99	cquiobp58 (OS-E/F)	1.98	cquiobp2 (OS-E/F)	1.98	agamobp14 (Mclassic9)	1.97	agamobp15 .16 (OS-E/F)	1.93	yes			repellent	4154948
240	benzaldehyde	C7H6O	aromatic	106	0	1	1	1.5	1	0	8	1.75±0.16	OS-E/F	1.92±0.22	cquiobp13 (OBP19a)	2.21	cquiobp58 (OS-E/F)	2.20	aaegobp56 (OBP19a)	2.11	cquiobp2 (OS-E/F)	2.07	agamobp6 (OBP19a)	2.05	yes			repellent	18306972
8785	benzyl acetate	C9H10O2	aromatic	150	0	2	3	2	1	0	11	1.87±0.19	OS-E/F	2.13±0.28	cquiobp58 (OS-E/F)	2.36	cquiobp13 (OBP19a)	2.30	cquiobp5 (OS-E/F)	2.25	cquiobp2 (OS-E/F)	2.24	agamobp15 .16 (OS-E/F)	2.23					
264	butanoic acid	C4H8O2	acid	88.1	1	2	2	0.8	1	0	6	1.52±0.25	mclassic2	1.77±0.46	cquiobp61 (Bombyx)	2.21	cquiobp28 (Pbprp2/5)	2.17	agamobp19 (OBP19a)	2.16	agamobp9 (OBP99a)	2.13	cquiobp29 (Pbprp2/5)	2.03	yes		yes	attractant	15703334
273	cadaverine	C5H14N2	amine	102	2	2	4	-0.6	1	0	7	2.00±0.43	mclassic7	2.63±0.64	cquiobp33 (Pbprp2/5)	3.12	cquiobp43 (OBP99a)	3.10	cquiobp31 (Pbprp2/5)	2.91	aaegobp17 (Mclassic7)	2.85	cquiobp34 (Pbprp2/5)	2.78				attractant	
3718401	cadaverine+2	C5H16N2+2	amine	104	2	0	4	-0.6	1	2	7	1.98±0.44	mclassic7	2.50±0.61	cquiobp31 (Pbprp2/5)	3.15	cquiobp33 (Pbprp2/5)	3.11	cquiobp43 (OBP99a)	2.96	cquiobp34 (Pbprp2/5)	2.89	agamobp29 (OBP59a)	2.87				attractant	
10364	carvacrol	C10H14O	terpene	150	1	1	1	3.1	9	0	11	2.01±0.19	OS-E/F	2.27±0.24	agamobp6 (OBP19a)	2.49	cquiobp58 (OS-E/F)	2.45	cquiobp13 (OBP19a)	2.42	cquiobp5 (OS-E/F)	2.41	cquiobp2 (OS-E/F)	2.40				repellent	
5281167	cis-3-hexen-1-ol	C6H10O	alcohol	98.1	0	1	3	1.5	1	0	7	1.63±0.16	OS-E/F	1.80±0.21	aaegobp56 (OBP19a)	2.05	cquiobp13 (OBP19a)	2.04	cquiobp58 (OS-E/F)	2.00	cquiobp2 (OS-E/F)	1.91	cquiobp5 (OS-E/F)	1.90					
7794	Citronellal	C10H18O	terpene	154	0	1	5	3	2	0	11	1.88±0.21	OS-E/F	2.21±0.27	cquiobp2 (OS-E/F)	2.44	cquiobp58 (OS-E/F)	2.39	aaegobp38 (OS-E/F)	2.37	agamobp6 (OBP19a)	2.32	cquiobp13 (OBP19a)	2.26			yes	repellent	8406635
325	cuminy alcohol	C10H14O	terpene	150	1	1	2	2.3	1	0	11	1.94±0.20	OS-E/F	2.24±0.29	cquiobp2 (OS-E/F)	2.51	cquiobp13 (OBP19a)	2.37	cquiobp4 (OS-E/F)	2.35	aaegobp38 (OS-E/F)	2.32	cquiobp58 (OS-E/F)	2.30				repellent	

¹Pubchem ID, ²Compound name, ³chemical formula, ⁴molecular weight, ⁵n° of hydrogen donor atoms, ⁶n° of hydrogen acceptor atoms, ⁷n° of rotatable bonds, ⁸partition coefficient, ⁹n° of tautomers, ¹⁰charge, ¹¹n° of heavy atoms, ¹²⁻¹⁴results of docking experiments, ¹²mean ligand efficiency for the ligand (SILE), ¹³cluster having best SILE value, ¹⁴average SILE value of best cluster, ¹⁵Presence in human emanation, ¹⁶Affect oviposition behavior, ¹⁷Affect other behaviors of mosquitoes, ¹⁸known activity of the compound, ¹⁹Pubmed ID or DOI number.

Pubchem ID ¹	Compound ²	Formula ³	Type	mw ⁴	H ⁵ don.	H ⁶ acc.	Rotat bonds ⁷	logp ⁸	Tau ⁹	Chrg ¹⁰	HA ¹¹	General mean SILE ¹²	Top cluster ¹³	avg of top cluster ¹⁴	Top #1	SILE	Top #2	SILE	Top #3	SILE	Top #4	SILE	Top #5	SILE	Hum. eman. ¹⁵	Ovip. ¹⁶	Bhv. ¹⁷	activity ¹⁸	PMID/DOI ¹⁹
7967	cyclohexanone	C6H10O	ketone	98.1	0	1	0	0.8	2	0	7	1.82±0.16	OS-E/F	2.01±0.19	cquiobp58 (OS-E/F)	2.36	cquiobp13 (OBP19a)	2.24	aaegobp56 (OBP19a)	2.15	cquiobp70 (Bombyx)	2.13	aaegobp18 (Mclassic9)	2.09	yes				10.1023/A:1005475422978
8175	decanal	C10H20O	aldehyde	156	0	1	8	3.8	2	0	11	1.71±0.21	OS-E/F	2.02±0.31	cquiobp2 (OS-E/F)	2.23	cquiobp58 (OS-E/F)	2.20	aaegobp38 (OS-E/F)	2.15	agamobp2 (OBP19a)	2.15	agamobp6 (OBP19a)	2.09	yes			repellent	18306972
2969	decanoic acid	C10H20O2	acid	172	1	2	8	4.1	1	0	12	1.70±0.24	mclassic2	1.96±0.36	agamobp14 (Mclassic9)	2.25	cquiobp58 (OS-E/F)	2.24	cquiobp61 (Bombyx)	2.21	aaegobp77 (Mclassic2)	2.21	cquiobp28 (Pbprp2/5)	2.19	yes		yes	attractant	10872864
4284	DEET	C12H17NO	amide	191	0	1	3	2	1	0	14	1.92±0.21	OS-E/F	2.18±0.29	cquiobp13 (OBP19a)	2.58	agamobp19 (OBP19a)	2.41	aaegobp78 (Bombyx)	2.40	agamobp6 (OBP19a)	2.37	cquiobp2 (OS-E/F)	2.34				repellent	12428949
12810	delta-decalactone	C10H18O2	lactone	170	0	2	4	2.5	2	0	12	2.06±0.21	OS-E/F	2.37±0.29	cquiobp2 (OS-E/F)	2.55	cquiobp58 (OS-E/F)	2.50	aaegobp38 (OS-E/F)	2.48	cquiobp4 (OS-E/F)	2.47	cquiobp13 (OBP19a)	2.42				repellent	
17076	DEPA	C12H17NO	amide	191	0	1	4	2.1	1	0	14	1.94±0.22	OS-E/F	2.24±0.28	cquiobp13 (OBP19a)	2.56	cquiobp2 (OS-E/F)	2.52	agamobp19 (OBP19a)	2.44	cquiobp58 (OS-E/F)	2.41	aaegobp78 (Bombyx)	2.35				repellent	
1068	dimethylsulfide	C2H6S	sulfur	62.1	0	0	0	0.9	1	0	3	1.18±0.09	Pbprp1	1.29±0.14	cquiobp27 (Pbprp2/5)	1.40	agamobp3 (OS-E/F)	1.36	cquiobp2 (OS-E/F)	1.36	aaegobp77 (Mclassic2)	1.36	agamobp7 (Pbprp1)	1.34	yes			attractant	6851177
3893	dodecanoic acid	C12H24O2	acid	200	1	2	10	4.2	1	0	14	1.75±0.24	mclassic2	2.00±0.34	cquiobp61 (Bombyx)	2.35	cquiobp58 (OS-E/F)	2.27	aaegobp77 (Mclassic2)	2.23	cquiobp4 (OS-E/F)	2.21	aaegobp2 (Pbprp1)	2.19	yes		yes	attractant	10872864
5281168	E2-hexenal	C6H10O	aldehyde	98.1	0	1	3	1.5	1	0	7	2.38±0.52	mclassic5	2.70±0.80	agamobp19 (OBP19a)	3.56	aaegobp78 (Bombyx)	3.46	cquiobp4 (OS-E/F)	3.46	cquiobp1 (OS-E/F)	3.39	cquiobp56 (Pbprp2/5)	3.26				attractant	
5318042	E2-hexenol	C6H12O	alcohol	100	1	1	3	1.4	1	0	7	1.63±0.17	OS-E/F	1.83±0.23	cquiobp13 (OBP19a)	2.06	cquiobp58 (OS-E/F)	2.06	aaegobp56 (OBP19a)	2.04	cquiobp2 (OS-E/F)	1.98	agamobp6 (OBP19a)	1.93					
702	ethanol	C2H6O	alcohol	46.1	1	1	0	-0.1	1	0	3	1.30±0.10	OS-E/F	1.39±0.12	cquiobp58 (OS-E/F)	1.55	cquiobp42 (Pbprp2/5)	1.54	cquiobp41 (Pbprp2/5)	1.50	aaegobp12 (Mclassic3)	1.50	cquiobp28 (Pbprp2/5)	1.49	yes				6445980
8857	ethyl acetate	C4H8O2	ester	88.1	0	2	2	0.7	1	0	6	1.44±0.12	OS-E/F	1.53±0.20	aaegobp56 (OBP19a)	1.70	cquiobp58 (OS-E/F)	1.68	cquiobp2 (OS-E/F)	1.67	agamobp65 (Nogroup)	1.66	aaegobp13 (Mclassic4)	1.66					
7762	ethyl butyrate	C6H12O2	ester	116	0	2	4	1.3	1	0	8	1.52±0.15	OS-E/F	1.68±0.24	cquiobp2 (OS-E/F)	1.89	cquiobp58 (OS-E/F)	1.83	cquiobp5 (OS-E/F)	1.82	aaegobp56 (OBP19a)	1.81	cquiobp67 (Bombyx)	1.78	yes			attractant	10.1023/A:1005475422978
8025	ethyl formate	C3H6O2	ester	74.1	0	2	2	0.5	1	0	5	1.28±0.13	Pbprp1	1.37±0.20	cquiobp62 (Bombyx)	1.64	cquiobp58 (OS-E/F)	1.58	agamobp19 (OBP19a)	1.54	cquiobp42 (Pbprp2/5)	1.52	aaegobp18 (Mclassic9)	1.52					
31265	ethyl hexanoate	C8H16O2	ester	144	0	2	6	2.4	1	0	10	1.62±0.18	OS-E/F	1.86±0.23	cquiobp58 (OS-E/F)	2.07	cquiobp13 (OBP19a)	2.03	cquiobp2 (OS-E/F)	2.01	cquiobp5 (OS-E/F)	2.00	aaegobp56 (OBP19a)	1.93	yes			attractant	10.1023/A:1005475422978
7749	ethyl propanoate	C5H10O2	ester	102	0	2	3	1.2	1	0	7	1.46±0.13	OS-E/F	1.59±0.17	cquiobp2 (OS-E/F)	1.83	cquiobp58 (OS-E/F)	1.78	aaegobp56 (OBP19a)	1.74	cquiobp5 (OS-E/F)	1.70	aaegobp79 (Nogroup)	1.67					
2758	Eucalyptol	C10H18O	terpene	154	0	1	0	2.5	1	0	11	1.96±0.20	OS-E/F	2.17±0.26	cquiobp13 (OBP19a)	2.52	agamobp6 (OBP19a)	2.50	aaegobp38 (OS-E/F)	2.36	aaegobp78 (Bombyx)	2.32	cquiobp68 (Bombyx)	2.29				repellent	
192578	eucamalol	C10H16O2	terpene	168	1	2	2	1.4	1	0	12	2.03±0.21	OS-E/F	2.32±0.26	cquiobp2 (OS-E/F)	2.63	cquiobp13 (OBP19a)	2.46	aaegobp38 (OS-E/F)	2.43	agamobp6 (OBP19a)	2.43	aaegobp56 (OBP19a)	2.38					
12813	gamma-decalactone	C10H18O2	lactone	170	0	2	5	2.7	2	0	12	2.00±0.21	OS-E/F	2.29±0.30	cquiobp2 (OS-E/F)	2.54	aaegobp38 (OS-E/F)	2.41	agamobp14 (Mclassic9)	2.37	cquiobp58 (OS-E/F)	2.35	aaegobp2 (Pbprp1)	2.33				repellent	
637566	Geraniol	C10H18O	terpene	154	1	1	4	2.9	1	0	11	1.91±0.21	OS-E/F	2.26±0.25	cquiobp2 (OS-E/F)	2.49	aaegobp38 (OS-E/F)	2.39	cquiobp58 (OS-E/F)	2.34	aaegobp78 (Bombyx)	2.29	agamobp2 (OS-E/F)	2.29	yes		yes	repellent	15474566
1549026	geranyl acetate	C12H20O2	terpene	196	0	2	6	3.5	1	0	14	2.02±0.22	OS-E/F	2.35±0.30	cquiobp2 (OS-E/F)	2.65	aaegobp78 (Bombyx)	2.49	cquiobp13 (OBP19a)	2.45	agamobp2 (OS-E/F)	2.44	aaegobp38 (OS-E/F)	2.42				repellent	20160092
1549778	geranyl acetone	C13H22O	ketone	194	0	1	6	3.7	3	0	14	2.10±0.24	OS-E/F	2.48±0.27	cquiobp2 (OS-E/F)	2.75	aaegobp38 (OS-E/F)	2.65	agamobp2 (OS-E/F)	2.55	cquiobp13 (OBP19a)	2.54	cquiobp4 (OS-E/F)	2.53	yes		yes	repellent	15474566
8130	heptanal	C7H14O	aldehyde	114	0	1	5	2.3	2	0	8	1.59±0.17	OS-E/F	1.80±0.26	cquiobp58 (OS-E/F)	2.03	cquiobp13 (OBP19a)	2.01	aaegobp56 (OBP19a)	1.98	cquiobp2 (OS-E/F)	1.96	cquiobp69 (Bombyx)	1.95	yes				12322940

¹Pubchem ID, ²Compound name, ³chemical formula, ⁴molecular weight, ⁵n° of hydrogen donor atoms, ⁶n° of hydrogen acceptor atoms, ⁷n° of rotatable bonds, ⁸partition coefficient, ⁹n° of tautomers, ¹⁰charge, ¹¹n° of heavy atoms, ¹²⁻¹⁴results of docking experiments, ¹²mean ligand efficiency for the ligand (SILE), ¹³cluster having best SILE value, ¹⁴average SILE value of best cluster, ¹⁵Presence in human emanation, ¹⁶Affect oviposition behavior, ¹⁷Affect other behaviors of mosquitoes, ¹⁸known activity of the compound, ¹⁹Pubmed ID or DOI number.

Pubchem ID ¹	Compound ²	Formula ³	Type	mw ⁴	H ⁵ don.	H ⁶ acc.	Rotat bonds ⁷	logp ⁸	Tau ⁹	Chrg ¹⁰	HA ¹¹	General mean SILE ¹²	Top cluster ¹³	avg of top cluster ¹⁴	Top #1	SILE	Top #2	SILE	Top #3	SILE	Top #4	SILE	Top #5	SILE	Hum. eman. ¹⁵	Ovip. ¹⁶	Bhv. ¹⁷	activity ¹⁸	PMID/DOI ¹⁹
8900	heptane	C7H16	alkane	100	0	0	4	4.4	1	0	7	1.56±0.16	OS-E/F	1.70±0.23	cquiobp13 (OBP19a)	1.93	cquiobp2 (OS-E/F)	1.93	cquiobp58 (OS-E/F)	1.92	cquiobp64 (Bombyx)	1.89	cquiobp5 (OS-E/F)	1.87	yes				12322940
8094	heptanoic acid	C7H14O2	acid	130	1	2	5	2.5	1	0	9	1.61±0.24	mclassic2	1.93±0.38	cquiobp61 (Bombyx)	2.24	aaegobp77 (Mclassic2)	2.23	agamobp9 (OBP99a)	2.19	cquiobp28 (Pbprp2/5)	2.18	cquiobp58 (OS-E/F)	2.12	yes		yes	attractant	10872864
985	hexadecanoic acid	C16H32O2	acid	256	1	2	14	6.4	1	0	18	1.80±0.24	OS-E/F	2.08±0.29	cquiobp4 (OS-E/F)	2.42	cquiobp37 (Pbprp2/5)	2.42	cquiobp58 (OS-E/F)	2.31	agamobp15 .16 (OS-E/F)	2.25	agamobp19 (OBP19a)	2.21	yes		yes	attractant	12322940
6184	hexanal	C6H12O	aldehyde	100	0	1	4	1.8	2	0	7	1.54±0.16	OS-E/F	1.71±0.22	cquiobp58 (OS-E/F)	1.94	aaegobp56 (OBP19a)	1.94	cquiobp13 (OBP19a)	1.89	cquiobp2 (OS-E/F)	1.85	cquiobp69 (Bombyx)	1.84	yes			attractant	20017925
8892	hexanoic acid	C6H12O2	acid	116	1	2	4	1.9	1	0	8	1.57±0.24	mclassic2	1.88±0.36	cquiobp28 (Pbprp2/5)	2.22	cquiobp61 (Bombyx)	2.13	agamobp9 (OBP99a)	2.11	aaegobp77 (Mclassic2)	2.09	agamobp24 (Mclassic2)	2.02	yes		yes	attractant	10872864
125098	Icaridin	C12H23NO3	heterocyclic	229	1	3	5	2	1	0	16	1.97±0.23	OS-E/F	2.26±0.33	cquiobp2 (OS-E/F)	2.66	cquiobp4 (OS-E/F)	2.57	aaegobp38 (OS-E/F)	2.46	cquiobp46 (Mclassic9)	2.45	cquiobp13 (OBP19a)	2.43				repellent	
798	indole (skatole)	C8H7N	aromatic	117	1	0	0	2.1	1	0	9	1.82±0.15	OS-E/F	1.98±0.21	cquiobp58 (OS-E/F)	2.27	cquiobp13 (OBP19a)	2.15	cquiobp2 (OS-E/F)	2.10	agamobp6 (OBP19a)	2.10	agamobp2 (OS-E/F)	2.02	yes	yes	yes	attractant	10945049
104150	IR3535	C11H21NO3	amino acid	215	0	3	8	1.2	1	0	15	1.70±0.23	OS-E/F	1.98±0.38	cquiobp2 (OS-E/F)	2.28	cquiobp4 (OS-E/F)	2.14	cquiobp46 (Mclassic9)	2.12	aaegobp77 (Mclassic2)	2.11	aaegobp38 (OS-E/F)	2.10				repellent	
31276	isoamyl acetate	C7H14O2	ester	130	0	2	4	2	1	0	9	1.66±0.17	OS-E/F	1.88±0.24	cquiobp58 (OS-E/F)	2.05	cquiobp13 (OBP19a)	2.04	agamobp65 (Nogroup)	2.02	cquiobp2 (OS-E/F)	1.99	aaegobp56 (OBP19a)	1.99					
8038	isobutyl acetate	C6H12O2	ester	116	0	2	3	1.8	1	0	8	1.63±0.15	OS-E/F	1.80±0.17	aaegobp56 (OBP19a)	2.02	cquiobp58 (OS-E/F)	2.00	cquiobp13 (OBP19a)	1.98	cquiobp2 (OS-E/F)	1.96	cquiobp67 (Bombyx)	1.92					
6590	isobutyric acid	C4H8O2	acid	88.1	1	2	1	0.8	1	0	6	1.57±0.25	mclassic2	1.81±0.46	cquiobp61 (Bombyx)	2.23	cquiobp28 (Pbprp2/5)	2.21	agamobp9 (OBP99a)	2.18	agamobp19 (OBP19a)	2.12	cquiobp29 (Pbprp2/5)	2.11	yes		yes	attractant	8887339
10430	isovaleric acid	C5H10O2	acid	102	1	2	2	1.2	1	0	7	1.59±0.24	mclassic2	1.87±0.38	cquiobp28 (Pbprp2/5)	2.24	cquiobp61 (Bombyx)	2.20	agamobp9 (OBP99a)	2.13	cquiobp29 (Pbprp2/5)	2.13	agamobp19 (OBP19a)	2.11	yes	yes	yes	attractant	19058627
107689	L(+) lactic acid	C3H6O3	acid	90.1	2	3	1	-0.7	1	0	6	1.94±0.28	mclassic2	2.14±0.52	agamobp19 (OBP19a)	2.78	aaegobp18 (Mclassic9)	2.60	agamobp9 (OBP99a)	2.57	cquiobp29 (Pbprp2/5)	2.51	cquiobp61 (Bombyx)	2.50	yes		yes	attractant	11583442
22311	limonene	C10H16	terpene	136	0	0	1	3.4	1	0	10	2.03±0.18	OS-E/F	2.26±0.24	cquiobp2 (OS-E/F)	2.48	cquiobp13 (OBP19a)	2.47	cquiobp58 (OS-E/F)	2.40	cquiobp5 (OS-E/F)	2.38	agamobp6 (OBP19a)	2.38				repellent	
22310	linalool oxide (furanoid)	C10H18O2	terpene	170	1	2	2	1.4	1	0	12	2.12±0.21	OS-E/F	2.36±0.27	cquiobp13 (OBP19a)	2.64	cquiobp4 (OS-E/F)	2.61	cquiobp2 (OS-E/F)	2.52	agamobp6 (OBP19a)	2.47	cquiobp58 (OS-E/F)	2.47	yes			repellent	10.1023/A: 100547542 2978
887	methanol	CH4O	alcohol	32	1	1	0	-0.5	1	0	2	1.25±0.10	Nogroup	1.35±0.18	agamobp12 (Mclassic9)	1.64	cquiobp30 (Pbprp2/5)	1.53	cquiobp41 (Pbprp2/5)	1.51	cquiobp73 (Bombyx)	1.51	cquiobp42 (Pbprp2/5)	1.51	yes				10.1023/A: 100547542 2978
7150	methyl benzoate	C8H8O2	aromatic	136	0	2	2	2.1	1	0	10	1.78±0.17	OS-E/F	2.00±0.21	cquiobp13 (OBP19a)	2.28	cquiobp58 (OS-E/F)	2.26	cquiobp2 (OS-E/F)	2.22	aaegobp38 (OS-E/F)	2.11	agamobp6 (OBP19a)	2.11				repellent	
8091	methyl caprylate	C9H18O2	ester	158	0	2	7	3.6	1	0	11	1.67±0.20	OS-E/F	1.95±0.29	cquiobp2 (OS-E/F)	2.19	cquiobp13 (OBP19a)	2.12	cquiobp58 (OS-E/F)	2.10	aaegobp38 (OS-E/F)	2.07	cquiobp5 (OS-E/F)	2.02					
11124	methyl propanoate	C4H8O2	ester	88.1	0	2	2	0.8	1	0	6	1.41±0.13	OS-E/F	1.51±0.20	aaegobp15 (Mclassic1)	1.82	cquiobp2 (OS-E/F)	1.70	cquiobp58 (OS-E/F)	1.68	aaegobp18 (Mclassic9)	1.67	aaegobp79 (Nogroup)	1.63					
4133	methyl salicylate	C8H8O3	aromatic	152	1	3	2	2.3	4	0	11	1.81±0.17	OS-E/F	2.03±0.19	cquiobp58 (OS-E/F)	2.36	aaegobp19 (Mclassic9)	2.22	cquiobp2 (OS-E/F)	2.19	cquiobp13 (OBP19a)	2.16	cquiobp57 (OBP19a)	2.14				repellent	
33094	methyl-2-methyl benzoate	C9H10O2	aromatic	150	0	2	2	2.8	1	0	11	1.84±0.19	OS-E/F	2.07±0.26	cquiobp58 (OS-E/F)	2.39	cquiobp13 (OBP19a)	2.34	aaegobp38 (OS-E/F)	2.24	agamobp6 (OBP19a)	2.24	cquiobp2 (OS-E/F)	2.22					
92770	Nepetalactone	C10H14O2	heterocyclic	166	0	2	0	1.9	2	0	12	2.11±0.20	OS-E/F	2.38±0.31	cquiobp2 (OS-E/F)	2.61	cquiobp58 (OS-E/F)	2.57	cquiobp5 (OS-E/F)	2.55	aaegobp38 (OS-E/F)	2.45	agamobp6 (OBP19a)	2.45				repellent	
31289	nonanal	C9H18O	aldehyde	142	0	1	7	3.3	2	0	10	1.68±0.21	OS-E/F	1.96±0.32	cquiobp58 (OS-E/F)	2.24	cquiobp2 (OS-E/F)	2.20	cquiobp69 (Bombyx)	2.15	aaegobp38 (OS-E/F)	2.06	agamobp14 (Mclassic9)	2.01	yes			repellent	18306972

¹Pubchem ID, ²Compound name, ³chemical formula, ⁴molecular weight, ⁵n° of hydrogen donor atoms, ⁶n° of hydrogen acceptor atoms, ⁷n° of rotatable bonds, ⁸partition coefficient, ⁹n° of tautomers, ¹⁰charge, ¹¹n° of heavy atoms, ¹²⁻¹⁴results of docking experiments, ¹²mean ligand efficiency for the ligand (SILE), ¹³cluster having best SILE value, ¹⁴average SILE value of best cluster, ¹⁵Presence in human emanation, ¹⁶Affect oviposition behavior, ¹⁷Affect other behaviors of mosquitoes, ¹⁸known activity of the compound, ¹⁹Pubmed ID or DOI number.

Pubchem ID ¹	Compound ²	Formula ³	Type	mw ⁴	H ⁵ don.	H ⁶ acc.	Rotat bonds ⁷	logp ⁸	Tau ⁹	Chrg ¹⁰	HA ¹¹	General mean SILE ¹²	Top cluster ¹³	avg of top cluster ¹⁴	Top #1	SILE	Top #2	SILE	Top #3	SILE	Top #4	SILE	Top #5	SILE	Hum. eman. ¹⁵	Ovip. ¹⁶	Bhv. ¹⁷	activity ¹⁸	PMID/DOI ¹⁹
8158	nonanoic acid	C9H18O2	acid	158	1	2	7	3.5	1	0	11	1.67±0.23	mclassic2	1.91±0.41	cquiobp58 (OS-E/F)	2.29	agamobp14 (Mclassic9)	2.26	aaegobp77 (Mclassic2)	2.26	cquiobp61 (Bombyx)	2.22	agamobp15 .16 (OS-E/F)	2.18	yes		yes	attractant	8887339
5281	octadecanoic acid	C18H36O2	acid	284	1	2	16	7.4	1	0	20	1.34±0.38	OBP99a	1.70±0.58	cquiobp5 (OS-E/F)	2.45	cquiobp13 (OBP19a)	2.36	agamobp1 (OS-E/F)	2.29	aaegobp36 (OS-E/F)	2.24	aaegobp37 (OS-E/F)	2.21	yes			attractant	
445639	octadecenoic acid	C18H34O2	acid	282	1	2	15	6.5	1	0	20	1.83±0.29	Pbprp1	2.10±0.46	agamobp15 .16 (OS-E/F)	2.43	cquiobp37 (Pbprp2/5)	2.43	cquiobp56 (Pbprp2/5)	2.38	cquiobp1 (OS-E/F)	2.36	agamobp19 (OBP19a)	2.32	yes			attractant	
454	octanal	C8H16O	aldehyde	128	0	1	6	2.7	2	0	9	1.64±0.19	OS-E/F	1.90±0.23	cquiobp58 (OS-E/F)	2.12	cquiobp2 (OS-E/F)	2.09	cquiobp13 (OBP19a)	2.01	aaegobp38 (OS-E/F)	1.98	cquiobp4 (OS-E/F)	1.98	yes			repellent	18306972
379	octanoic acid	C8H16O2	acid	144	1	2	6	3	1	0	10	1.63±0.24	mclassic2	1.96±0.38	aaegobp77 (Mclassic2)	2.26	agamobp14 (Mclassic9)	2.25	agamobp9 (OBP99a)	2.23	cquiobp58 (OS-E/F)	2.19	cquiobp61 (Bombyx)	2.17	yes		yes	attractant	8887339
7991	pentanoic acid	C5H10O2	acid	102	1	2	3	1.4	1	0	7	1.53±0.25	mclassic2	1.82±0.41	cquiobp61 (Bombyx)	2.21	cquiobp28 (Pbprp2/5)	2.20	agamobp19 (OBP19a)	2.19	agamobp9 (OBP99a)	2.09	aaegobp77 (Mclassic2)	2.03	yes		yes	attractant	8887339
40326	permethrin	C21H20Cl2O3	heterocyclic	391	0	3	7	6.5	1	0	25	2.62±0.32	Pbprp1	3.02±0.50	agamobp19 (OBP19a)	3.41	cquiobp68 (Bombyx)	3.22	agamobp2 (OS-E/F)	3.14	aaegobp2 (Pbprp1)	3.13	cquiobp4 (OS-E/F)	3.07				repellent	
7654	phenethyl acetate	C10H12O2	aromatic	164	0	2	4	2.3	1	0	12	1.91±0.20	OS-E/F	2.20±0.31	cquiobp2 (OS-E/F)	2.40	cquiobp58 (OS-E/F)	2.36	aaegobp38 (OS-E/F)	2.32	cquiobp13 (OBP19a)	2.28	agamobp65 (Nogroup)	2.28					
996	phenol	C6H6O	aromatic	94.1	1	1	0	1.5	2	0	7	1.69±0.15	OS-E/F	1.88±0.19	cquiobp58 (OS-E/F)	2.25	cquiobp13 (OBP19a)	2.03	aaegobp18 (Mclassic9)	2.01	cquiobp2 (OS-E/F)	2.01	agamobp6 (OBP19a)	1.97	yes	yes			1583482
19100	PMD	C10H20O2	terpene	172	2	2	1	2.2	1	0	12	2.12±0.22	OS-E/F	2.40±0.30	cquiobp5 (OS-E/F)	2.66	agamobp6 (OBP19a)	2.66	aaegobp78 (Bombyx)	2.56	cquiobp13 (OBP19a)	2.56	cquiobp2 (OS-E/F)	2.55				repellent	
1032	propanoic acid	C3H6O2	acid	74.1	1	2	1	0.3	1	0	5	1.52±0.25	mclassic2	1.73±0.50	agamobp19 (OBP19a)	2.18	cquiobp61 (Bombyx)	2.17	agamobp9 (OBP99a)	2.16	cquiobp28 (Pbprp2/5)	2.15	cquiobp29 (Pbprp2/5)	2.02	yes		yes	attractant	8887339
7997	propyl acetate	C5H10O2	ester	102	0	2	3	1.2	1	0	7	1.50±0.14	OS-E/F	1.63±0.18	aaegobp56 (OBP19a)	1.84	cquiobp58 (OS-E/F)	1.82	cquiobp70 (Bombyx)	1.79	cquiobp2 (OS-E/F)	1.75	agamobp65 (Nogroup)	1.75					
1045	putrescine	C4H12N2	amine	88.2	2	2	3	-0.9	1	0	6	2.04±0.44	mclassic7	2.73±0.63	cquiobp33 (Pbprp2/5)	3.19	cquiobp43 (OBP99a)	3.18	cquiobp31 (Pbprp2/5)	3.09	aaegobp17 (Mclassic7)	3.02	agamobp29 (OBP59a)	2.93				attractant	
11005	tetradecanoic acid	C14H28O2	acid	228	1	2	12	5.3	1	0	16	1.79±0.26	Pbprp1	2.07±0.39	aaegobp2 (Pbprp1)	2.44	cquiobp4 (OS-E/F)	2.36	aaegobp77 (Mclassic2)	2.33	cquiobp77 (Mclassic2)	2.26	cquiobp37 (Pbprp2/5)	2.25	yes		yes	attractant	8887339
9256	thiazole	C3H3NS	heterocyclic	85.1	0	0	0	0.4	1	0	5	1.38±0.12	OS-E/F	1.50±0.19	cquiobp54 (Pbprp2/5)	1.70	cquiobp58 (OS-E/F)	1.68	cquiobp70 (Bombyx)	1.67	cquiobp29 (Pbprp2/5)	1.65	cquiobp2 (OS-E/F)	1.65					
12530	tridecanoic acid	C13H26O2	acid	214	1	2	11	4.7	1	0	15	1.78±0.26	OS-E/F	2.02±0.46	cquiobp58 (OS-E/F)	2.46	aaegobp77 (Mclassic2)	2.32	cquiobp61 (Bombyx)	2.31	aaegobp2 (Pbprp1)	2.26	agamobp24 (Mclassic2)	2.25	yes			attractant	10872864
5324489	Z2-hexenol	C6H12O	alcohol	100	1	1	3	1.4	1	0	7	1.62±0.17	OS-E/F	1.84±0.27	cquiobp58 (OS-E/F)	2.15	cquiobp13 (OBP19a)	2.05	aaegobp56 (OBP19a)	1.99	cquiobp5 (OS-E/F)	1.95	cquiobp2 (OS-E/F)	1.93					

¹Pubchem ID, ²Compound name, ³chemical formula, ⁴molecular weight, ⁵n° of hydrogen donor atoms, ⁶n° of hydrogen acceptor atoms, ⁷n° of rotatable bonds, ⁸partition coefficient, ⁹n° of tautomers, ¹⁰charge, ¹¹n° of heavy atoms, ¹²⁻¹⁴results of docking experiments, ¹²mean ligand efficiency for the ligand (SILE), ¹³cluster having best SILE value, ¹⁴average SILE value of best cluster, ¹⁵Presence in human emanation, ¹⁶Affect oviposition behavior, ¹⁷Affect other behaviors of mosquitoes, ¹⁸known activity of the compound, ¹⁹Pubmed ID or DOI number.

Table 6.2. Description of the various parameters of docking used in the optimization protocol based on the ability of AUTODOCK to reproduce the bound complex of the CquiOBP1-3OG complex for the large scale docking.

Run	Dimension of the box			Grid centre			Genetic algorithm parameters			Time	Machine
	X	Y	Z	x center	y center	z center	Num GA Runs	Pop size	Max num of evals		
1	58	44	56	29.156	37	41.01	100	150	2500000	1h 08m 45.98s	Valhalla
2	74	50	44	28.269	38.5	40.345	100	150	2500000	1h 11m 18.66s	titan1
3	74	50	58	29.269	38.5	40.345	100	150	2500000	1h 11m 15.28s	titan1
4	74	58	58	29.269	39.5	35.01	100	150	2500000	57m 42.85s	bioch-chd189.univ. run
5	74	58	58	29.269	39.5	35.01	100	300	2500000	1h 09m 47.05s	bioch-chd189.univ. run
6	74	58	58	29.269	39.5	35.01	100	150	25000000	9h 49m 21.86s	bioch-chd189.univ. run
7	70	60	72	31.156	39.673	38.01	100	300	25000000	10h 45m 34.25s	bioch-chd189.univ. run
8	70	62	72	31.156	38.673	38.01	100	300	25000000	10h 47m 27.05s	bioch-chd189.univ. run
9	70	54	66	31.156	38.673	38.01	100	150	2500000	1h 11m 07.35s	titan1
10	70	54	66	31.156	38.673	38.01	1000	150	2500000	error	titan1
11	70	54	66	31.156	38.673	38.01	100	150	25000000	11h 51m 48.01s	titan1
12	70	54	66	31.156	38.673	38.01	100	300	25000000	11h 58m 08.97s	titan1
13	70	54	66	31.156	38.673	38.01	100	150	10000000	4h 48m 23.97s	titan1
14	70	54	66	31.156	38.673	38.01	100	300	2500000	1h 13m 12.03s	titan1
15	70	54	66	31.156	38.673	38.01	256	150	2500000	3h 05m 15.16s	titan1

Table 6.3. SILE values derived from AUTODOCK energy values for ligands proved to experimentally bind to OBPs

Receptor Vs Ligand	Agam OBP1/17	Aedes OBP1/39	Cqui OBP1	Agam OBP4	Agam OBP3	Agam OBP19
Octanal	1.66	2.34	1.77	-	-	-
Nonanal	1.76	2.44	1.9	7.6	1.92	
Geranyl acetone	2.39	2.96	2.4	-	-	-
Decanal	1.83	2.5	1.97	1.52	-	-
DEET	2.1	2.7	2.07	-	-	-
3-methyl indole	2.06	2.69	2.00	-	-	-
indole	1.92	2.56	1.92	-	-	-
1-dodecanol	-	-	-	1.64	-	-
octanal	-	-	-	1.50	-	-
citronellal	2.02	-	-	1.76	2.18	2.17

7

Conclusion

The history of the involvement of mosquitoes in the transmission of infectious diseases dates back to 1902 with the work of Ronald Ross which was awarded that year's Nobel prize. In his paper titled "The Role of the Mosquito in the Evolution of the Malaria Parasite: The Recent Researches of Surgeon-Major Ronald Ross, I.M.S." he states "The practical applications of the discovery are immeasurable and the establishment of the fact that as the bite of the snake or the rabid dog inoculates the blood of the victims of these creatures so the mosquito conveys malaria, would open up a new and hopeful phase as regards the prevention of disease in the tropics" (Ronald Ross. 1898). Beginning with that for almost a century, researchers have been trying to divert mosquitoes from their pursuit of human blood. The field blossomed in the 1950s, when dozens of entomologists in several countries set out to discover what attracts females—the only mosquitoes that bite—to their hosts. However by the mid 1960's, most research on host attraction had stopped with the discovery of DDT. Later the development of resistant in mosquitoes to DDT and the other drawbacks of the use of DDT encouraged the need of other insecticides and the research gained momentum again. A large group of researchers focused on the components of sweat that play an important role in the host seeking process based on behavioral studies reviewed in Foster. (1995); Takken & Knols. (1999). In parallel the molecular dissection of the olfactory response in vertebrates and invertebrates had taken its shape revealing its complexity and the major molecules involved in reception and signal transduction. Analogous to the vertebrate olfactory system, the detection of odor molecules by insects involves odorant binding proteins (OBPs) and pheromone binding proteins (PBP). These proteins are believed to carry the compounds from the porous cuticular surface of the antennal sensilla through the sensillum lymph to the G-protein-coupled odorant receptors residing on the dendritic membrane of the olfactory sensory neurons. A nice breakthrough in the quest of recognizing the molecules involved in the odor reception of mosquitoes was brought by Catherine et al. (2002) with the discovery of 79 odor-receptor candidates only five of which had been known before. A number of studies focussed on the

functional aspects of these receptors with the single motivation of identifying compounds that interfere the mosquito host intersection and are still in progress.

In contrast to the odorant receptors, odorant and pheromone binding proteins (OBPs and PBPs) are found to be abundant in insect antennae. OBPs were first discovered in moths (Vogt and Riddiford. 1981) and have subsequently been found in a variety of insects. Beissmann et al. (2002) and Ishida et al. (2002) isolated the first OBPs in the mosquito genomes *Anopheles gambiae* and *Culex quinquefasciatus* the same year as the odorant receptors were reported. This speeded up the identification of a number of OBPs in the mosquito species subsequently reported in (Vogt et al. 2002; Zhou et al. 2004; Xu et al. 2003; Zhou et al. 2008; Pelletier et al. 2009; Armbruster et al. 2009; Viera and Rozas. 2011). A massive expansion to the currently known OBPs in the mosquito genomes is provided in this study. This expansion, mainly driven by gene duplicates, that is observed more specifically in the *Culicidae* species, brings new insights into the molecular basis for understanding the diversity of behavioral patterns adapted by the mosquitoes in response to varied and demanding ecological constraints. It sustains the hypothesis that genes involved in olfaction contribute to a gene expansion in mosquitoes (Arensburger et al. 2010), which exhibited a fast genome evolution with respect to ecological constraints. This highlights the probable complexity of the mechanism underlying olfaction and in particular, the probable combinatorial nature of odorant recognition. However and interestingly, though a massive expansion of the genes is described here, it is observed that these genes are still confined to three subfamilies which helps to reduce the complexity and eases the comprehension of the functional properties of these OBPs.

This work also provides a rational background for the naming of OBPs in the mosquito genomes which demands stabilization. It highlights that it is critical to have a consensus naming convention for mosquito OBPs.

The identification and detailed characterization of two-domain OBPs in this study emerges to be a major step in annotating the current knowledge about mosquito OBPs. The *Atypical* OBPs named after the presence of an uncharacterized long C-terminal end observed in these proteins are confirmed in this study to be indeed two domain OBPs. This study, also for the first time, establishes the origin of these OBPs. The *Atypical* OBPs appear to have a distant origin when compared to the *Classic* OBPs in terms of sequence conservations, but however they still hold a significant relation to the *Classic* OBPs. It can be speculated that either the two domains of OBPs in the case of these two domain proteins split to form two classic OBPs or two classic OBPs fused to form the two domain OBPs. More detailed phylogenetic analysis could provide more insights into

this aspect of analysis. Functional implications that demand this kind of adaptations could provide more clues for the identification of more potential targets for repellent molecules.

Conserved cysteines are confirmed to be the hallmark of OBPs in the arthropod OBP domains and it stands as a key feature to recognize OBP genes. All members of the *Classic*, *Plus C* and *Atypical* OBPs are found to retain this universal footprint of the OBPs. The only exception to this, is seen in the case of proteins called the *Minus C* OBPs that lack the second and fifth cysteines that are engaged in the formation of a disulphide bond. An apparition of this new form of OBPs in the *Culicidae* species is described for the first time, which is otherwise observed only in distantly related species (*Bombyx mori*). This stands as one of the groups among the *Classic* OBPs which show the highest observed sequence divergence. The *Minus C* feature however stands out to be an evolutionary adaptation specific to *Culicidae*s more precisely to the *Culex* species. This brings new questions on the evolutionary aspects of these species and their adaptation. A new type of *Atypical* OBPs which miss out some of the hallmark cysteines have also been observed and it is again specific to the *Culicidae* species. Members of *Anopheles* closely related to these proteins have the 12 cysteines and it is also observed that this cluster (*matype2*) is largely dominated by members from *Aedes* and *Culex* which lack the second and fifth cysteine in the first domain and four other cysteines (C1,C3,C4 and C6) in the second domain. This further confirms the functional adaptation specific to these two species. This specialization is further sustained by the fact that no *Classic* OBPs are detected as remote homologues of these genes. Interestingly one classic OBP in another subclass (*mclassic6*) is found to lack C2 and C5 (AegOBP76). Likewise another subclass of OBPs AegOBP77 in *mclassic2* also shows the absence of these cysteines. Thus *Minus C* proteins are an outstanding feature of the *Culicidae*s and their appearance can be attributed to their specific evolutionary dynamics.

In general, as described in Chapter 2, this work provides a very exhaustive and robust formalization of the sub-grouping of the genes in the different subfamilies. A detailed representation of this data is provided in the form of cluster specific sequence conservation patterns. This was rendered possible with the use of structure based alignments to infer the phylogenetic subgroups which would have been missed otherwise. Automated detection of OBP subtypes can be now achieved using these profiles. The use of structure analysis and their importance in aiding a robust classification pattern for the arthropod OBPs is described for the first time in this study and was not found to prevail in this field of analysis.

Based on this key feature of conservation of cysteine patterns, a method has been developed for the classification of the OBPs into the different subfamilies described in Chapter 3. The

accuracy of prediction observed in this method further confirms that the conservation of cysteines are indeed one of the best choices in the identification and classification of a diverse family of proteins which in general pose lot of challenges on regular identification and classification algorithms.

The conservation of cysteines have been found to contribute in maintaining the overall fold among a family of proteins inspite of high sequence divergence (Thangudu et al. 2005). Taking advantage of the overall conservation of cysteines that is observed among the *Classic* OBPs and as it has been established that the evolutionary pressure posed by the cysteine conservation can produce good models albeit the limitation of methods, a large scale modeling of these OBPs has been described in Chapter 4. These protein models are indeed relevant alternatives when there is lack of experimental data owing to the high conservation of the fold imposed by the cysteine conservation pattern. They provide the basis for the structural and functional characterization of the OBPs in the mosquitoes that have been explored in Chapter 6. This was further made affirmative by the comparison of the models obtained in this analysis with subsequently published crystal structures of AgamOBP4 and CquiOBP1 which confirmed the quality of the predicted models.

It is evident from the genomic characterization of the OBP repertoire that there are a number of new questions to be answered on the functional aspects of the OBPs. There are two aspects of unction that is addressed in this thesis: (i) the diversity of the odorant molecules that these OBPs can handle facilitating their recognition by the olfactory system and (ii) the mechanism by which an OBP could participate in the sensing of a particular odorant. Both these issues have been addressed by conducting simulation experiments described in Chapter 5 and Chapter 6.

Protein ligand binding experiments using the molecular docking approach was used towards the first objective which was to explore the ligand binding profiles of the different set of proteins and their corresponding clusters. A genome wide analysis towards the prediction of the ligand binding properties of *Classic* OBPs in mosquitoes is described for the first time in Chapter 6. The large scale data was preliminarily filtered for significant interactions by relying on new sets of objective criteria described in the literature to assess the binding efficiency of ligands independent of the size. This analysis has provided a massive aid in shaping the putative binding specificity profiles for all the known classic OBPs in mosquitoes. The results indicate the existence of subgroups of OBPs that potentially have a brand range spectrum towards odorant recognition (OSE/OSF, Pbprp1, OBP19a, *mclassic1-6*) and subgroups that indicate a narrow spectrum towards the recognition of odorants (Pbprp2/Pbprp5, Lush, Pbprp1, *mclassic7*, *mclassic8*, and OBP members closely related to *Bombyx mori* minus C proteins and *Drosophila* minus C proteins) suggesting a

combinatorial mode of recognition of odorants. Furthermore, the recent finding that the binding of a ligand like indole to AgamOBP4 is a necessary step towards the heterodimerization of the protein with another OBP further broadens the functional scope of ligand binding events and reinforces the idea of a combinatorial nature of odorant recognition by mosquito OBPs. These results are however yet to be viewed as a preliminary step toward in-depth functional characterization of the *Classic* OBPs in mosquitoes. The large dataset that was generated in this study has yet to be mined and analyzed on an individual basis to further identify the key structural features (residues involved, localization in the binding pocket...) that are involved in the recognition of the ligand for the protein ligand complexes that are predicted to be significant. It can be anticipated that this will provide insight on the observed sequence conservation patterns in the different clusters and their relation to the functional aspects of that particular cluster. It further provides a rational and a wealth of information for further experimental characterization in terms of ligand binding experiments. The validity of the predicted ligand binding studies is further exemplified with the accuracy of the predicted binding to the last two ligand bound crystal structures that were published, *i.e* AgamOBP1 with DEET and AgamOBP4 with indole, though these experiments were not carried out at identical pH conditions.

Indeed pH seems to play an important role in the mechanism of binding of OBPs to odorant molecules. This has been documented a few times in the literature with respect to OBPs from *Bombyx mori* and *Antheraea polyphemus*. Evidence of a somewhat similar mechanism in mosquitoes is described in Chapter 5 of this thesis through the molecular simulations of CquiOBP1-MOP complex (PDBID: 3OGN) in different pH conditions. The results indicate that pH changes might mediate conformational changes that are directed toward ligand delivery. A set of well characterized changes involving residues that participate in changing the orientation of a functional loop between helix3 and helix4 is described in detail. The essential dynamic analysis and the observed concerted disruption of key interactions compensated by new sets of interactions, confirm the flip of this important loop when the pH is lowered. The concomitant (i) closure of what is believed to be the entrance of the binding pocket, (ii) expansion of what could be an exit site of the ligand and (iii) migration of the ligand towards the putative exit site provides insights into the probable mechanism of how OBPs might deliver a ligand to the membrane bound receptors when it approaches the lowered pH environment surrounding them. The fact that this loop importantly accommodates charged residues that are highly conserved across all OBPs sustains the hypothesis that the flip of this loop could be a conserved mechanism observed among the OBPs. To what

extent this might participate in activating the membrane bound ORs as a bound complex as it is described in LUSH OBPs is yet to be explored.

Thus, the work provided in this thesis stands to be an extensive characterization of the OBPs in the *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* mosquito genomes in various dimensions starting from a genome based analysis to structure and function determination. Apart from providing interesting results, every part of this thesis has raised many interesting questions on the various aspects of OBPs which is put forth to the scientific community.

In particular, this work can readily be further expanded towards the structural and functional characterization of *PlusC* and *Atypical* proteins. Indeed, now that the *PlusC* fold is established experimentally (PDB:3PM2), comparative modelling could be used to model the different members of this subfamily. Also, as it has been established in this work that *Atypical* OBPs are indeed composed of two OBP domains each of which are distantly related to *Classic* OBPs. The modelling of each domain can follow the same methodology used here but further protein-protein docking would be required to investigate the precise interface between the two domains. Knowledge from crystallographic dimers might provide more interesting overviews in this line. The structural characterization of these *Atypical* proteins would definitely be a landmark towards the understanding of their functional implications in olfaction of mosquitoes and how they shape up the evolutionary dynamics of their olfactory system.

Also, the exact mechanism(s) by which OBPs are involved in the perception of odorant molecules by the olfactory system has yet to be established. Several theoretical models can be proposed (Figure 7.1) : (i) OBP monomers act as transporters for the odorant molecules and directly delivers them to the ORs, whereby signal firing is onset upon formation of OR-ligand complex, (ii) OBP monomers act as transporters for the odorant molecules and directly activate the ORs through formation of an OBP-OR complex at the membrane surface, (iii) OBP monomers act as transporters for the odorant molecules and indirectly activate the ORs through the activation of an intermediary membrane protein likewise the mechanism hypothesized for Lush OBP, (iv) upon binding of a ligand to an OBP monomer, heterodimerization is facilitated with another unbound (apo-) or bound OBP and this heterodimer complex would either deliver the ligand to the ORs or activates the ORs directly or indirectly.

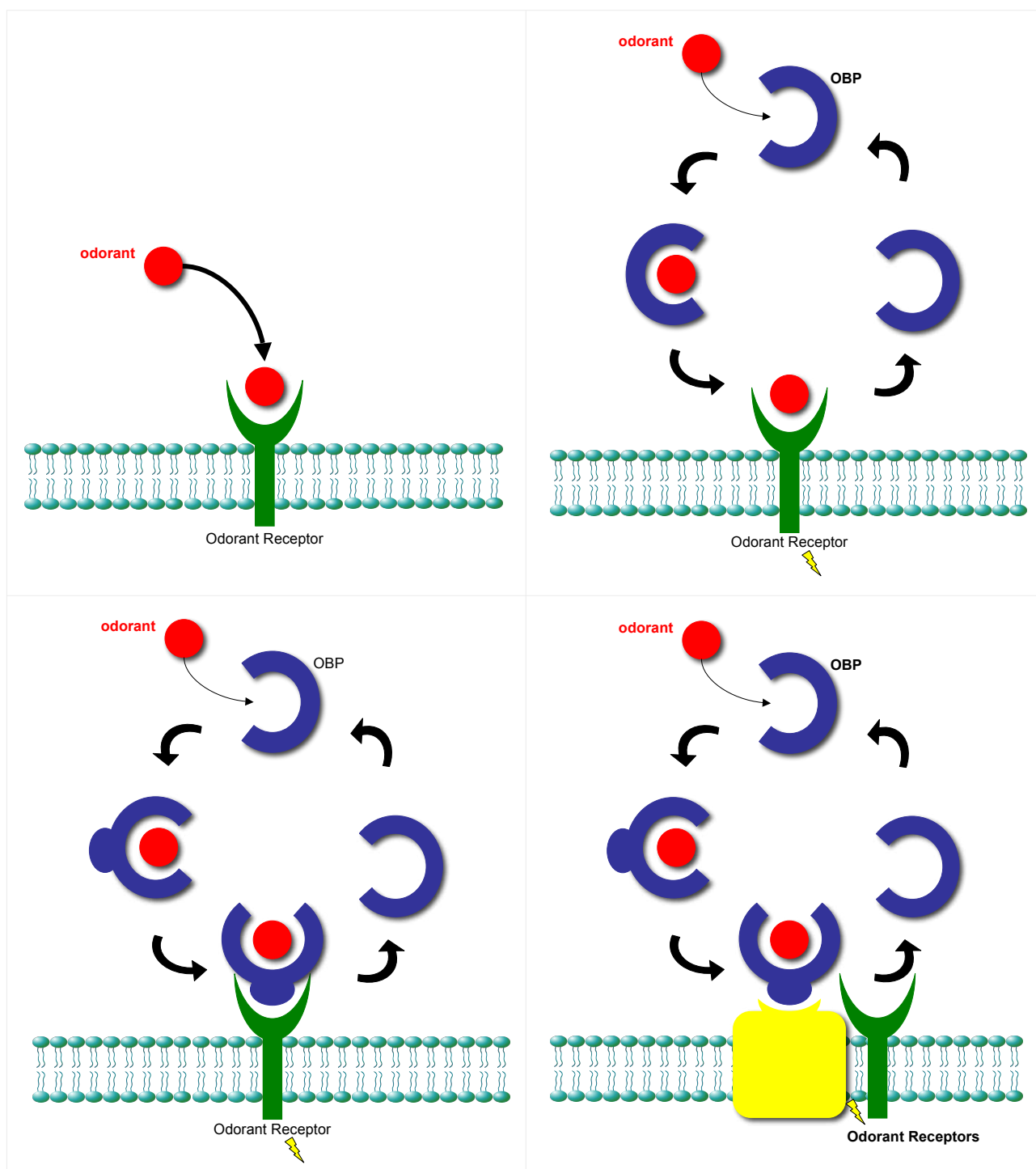


Figure 7.1. Representation of the various hypotheses about the molecular mechanism by which OBPs would be involved in olfaction in the mosquitoes. (a) describes the activation of the ORs by the odorant molecules without the requirement of an OBP (b) describes the delivery of the odorants by the OBPs to the olfactory receptors, (c) describes the direct activation of the ORs by an OBP-odorant complex, (d) describes the indirect activation of the ORs through the binding of the OBP-odorant complex to an accessory protein (continued on next page).

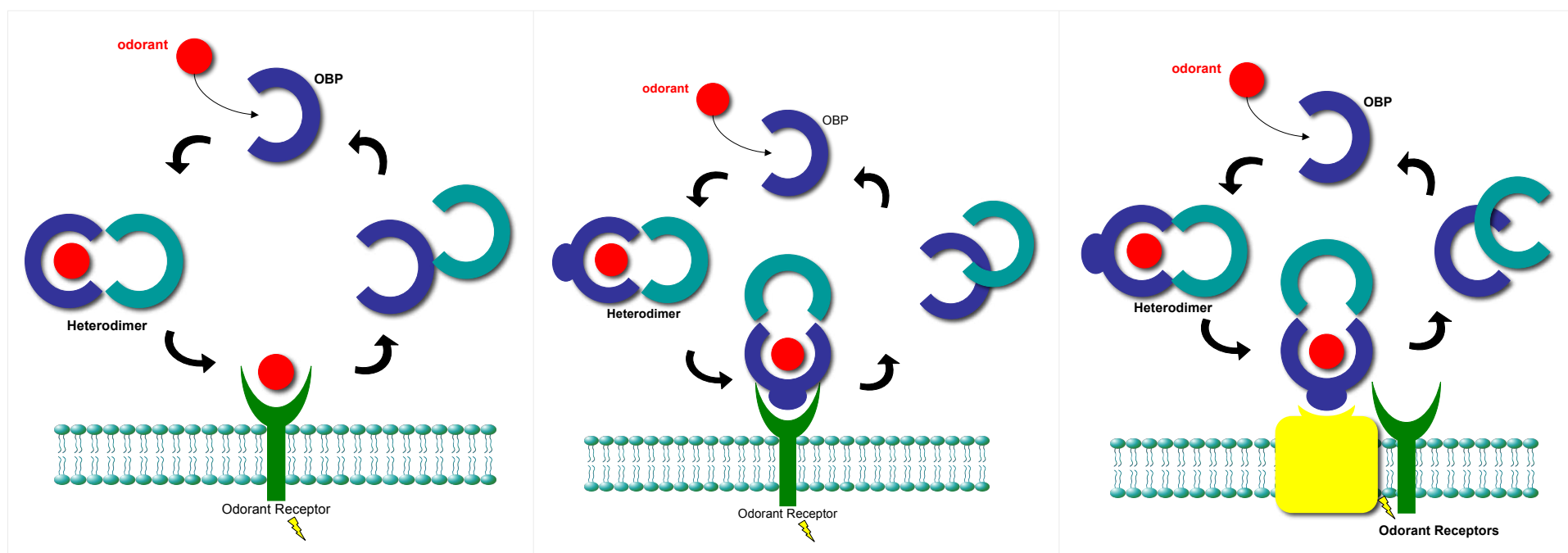


Figure 7.1 (contd). Cartoonic representation of the various hypothesis that can be derived for the mechanism of olfaction in the mosquitoes involving the role of OBPs. (e) describes the heterodimerization of the OBPs following the ligand binding to deliver the ligand to the ORs, (f) describes the direct activation of the ORs by the binding of the OBP-odorant-OBP heterodimeric complex to the ORs (g) describes the indirect activation of the ORs by the binding of the OBP-odorant-OBP heterodimeric complex to an accessory protein to the ORs (continued on the next page).

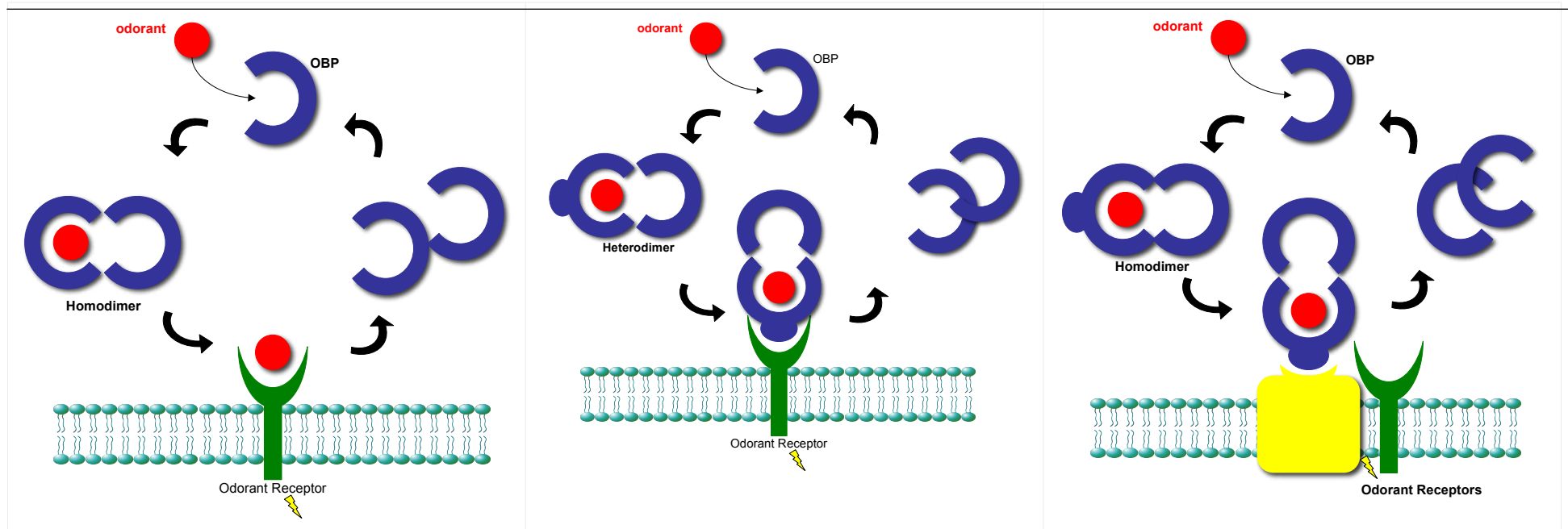


Figure 7.2 (contd) Cartoonic representation of the various hypothesis that can be derived for the mechanism of olfaction in the mosquitoes involving the role of OBPs. (h) describes the homodimerization of the OBPs following the ligand binding to deliver the ligand to the ORs, (i) describes the direct activation of the ORS by the binding of the OBP-odorant-OBP homodimeric complex to the ORs (j) describes the indirect activation of the ORS by the binding of the OBP-odorant-OBP homodimeric complex to an accessory protein to the ORs.

8

References

- Akutsu, T., Sekiguchi, K., Ohmori, T., and Sakurada, K. 2006. Individual comparisons of the levels of (E)-3-methyl-2-hexenoic acid, an axillary odor-related compound, in Japanese. *Chemical senses* **31**: 557-563.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**: 3389-3402.
- Apostolico, A. and Giancarlo, R. 1998. Sequence alignment in molecular biology. *J Comput Biol* **5**: 173-196.
- Arensburger, P., Megy, K., Waterhouse, R.M., Abrudan, J., Amedeo, P., Antelo, B., Bartholomay, L., Bidwell, S., Caler, E., Camara, F. et al. 2010. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. *Science (New York, N.Y)* **330**: 86-88.
- Barnard, D.R. 2005. Biological assay methods for mosquito repellents. *Journal of the American Mosquito Control Association* **21**: 12-16.
- Bartelt, R., AM, S., and LL, J. 1985. cis-vaccenyl Acetate as an Aggregation Pheromone in *Drosophila melanogaster*. *Journal of chemical ecology* **11**: 1747-1756.
- Benham, C.J. and Jafri, M.S. 1993. Disulfide bonding patterns and protein topologies. *Protein Sci* **2**: 41-54.
- Bentley, M.D., McDaniel, I.N., and Davis, E.E. 1982. Studies of 4-methylcyclohexanol: an *Aedes triseriatus* (Diptera: Culicidae) oviposition attractant. *Journal of medical entomology* **19**: 589-592.
- Benton, R., Vannice, K.S., and Vosshall, L.B. 2007. An essential role for a CD36-related receptor in pheromone detection in *Drosophila*. *Nature* **450**: 289-293.
- Bernier, U.R., Kline, D.L., Barnard, D.R., Schreck, C.E., and Yost, R.A. 2000. Analysis of human skin emanations by gas chromatography/mass spectrometry. 2. Identification of volatile compounds that are candidate attractants for the yellow fever mosquito (*Aedes aegypti*). *Analytical chemistry* **72**: 747-756.
- Bernier, U.R., Kline, D.L., Schreck, C.E., Yost, R.A., and Barnard, D.R. 2002. Chemical analysis of human skin emanations: comparison of volatiles from humans that differ in attraction of *Aedes aegypti* (Diptera: Culicidae). *Journal of the American Mosquito Control Association* **18**: 186-195.
- Bette, S., Breer, H., and Krieger, J. 2002. Probing a pheromone binding protein of the silkworm *Antheraea polyphemus* by endogenous tryptophan fluorescence. *Insect biochemistry and molecular biology* **32**: 241-246.
- Biessmann, H., Andronopoulou, E., Biessmann, M.R., Douris, V., Dimitratos, S.D., Eliopoulos, E., Guerin, P.M., Iatrou, K., Justice, R.W., Krober, T. et al. 2010. The *Anopheles gambiae*

- odorant binding protein 1 (AgamOBP1) mediates indole recognition in the antennae of female mosquitoes. *PLoS one* **5**: e9471.
- Bissantz, C., Folkers, G., and Rognan, D. 2000. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of medicinal chemistry* **43**: 4759-4767.
- Blackwell, A. and Johnson, S.N. 2000. Electrophysiological investigation of larval water and potential oviposition chemo-attractants for *Anopheles gambiae* s.s. *Annals of tropical medicine and parasitology* **94**: 389-398.
- Blake, J.D. and Cohen, F.E. 2001. Pairwise sequence alignment below the twilight zone. *Journal of molecular biology* **307**: 721-735.
- Blundell, T.L., Sibanda, B.L., Sternberg, M.J., and Thornton, J.M. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**: 347-352.
- Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science (New York, N.Y.)* **253**: 164-170.
- Braks, M.A.H., Meijerink, J., and Takken, W. 2001. The response of the malaria mosquito, *Anopheles gambiae*, to two components of human sweat, ammonia and l-lactic acid, in an olfactometer. *Physiological Entomology* **26**: 142-148.
- Breer, H., Krieger, J., and Raming, K. 1990. A novel class of binding proteins in the antennae of silk moth *Antheraea Pernyi*. *Insect Biochemistry* **20**: 735-740.
- Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. et al. 2009. CHARMM: the biomolecular simulation program. *Journal of computational chemistry* **30**: 1545-1614.
- Brooksbank, B.W., Brown, R., and Gustafsson, J.A. 1974. The detection of 5 α -androst-16-en-3 α -ol in human male axillary sweat. *Experientia* **30**: 864-865.
- Brown, A.W.A. 1966. The attraction of Mosquitoes to Hosts. *JAMA: The Journal of the American Medical Association* **196**: 249-252.
- Browne, W.J., North, A.C., Phillips, D.C., Brew, K., Vanaman, T.C., and Hill, R.L. 1969. A possible three-dimensional structure of bovine α -lactalbumin based on that of hen's egg-white lysozyme. *Journal of molecular biology* **42**: 65-86.
- Butte, A.J. 2001. Challenges in bioinformatics: infrastructure, models and analytics. *Trends in biotechnology* **19**: 159-160.
- Bystroff, C. and Baker, D. 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *Journal of molecular biology* **281**: 565-577.
- Calvo, E., deBianchi, A.G., James, A.A., and Marinotti, O. 2002. The major acid soluble proteins of adult female *Anopheles darlingi* salivary glands include a member of the D7-related family of proteins. *Insect biochemistry and molecular biology* **32**: 1419-1427.
- Calvo, E., Mans, B.J., Andersen, J.F., and Ribeiro, J.M. 2006. Function and evolution of a mosquito salivary protein family. *The Journal of biological chemistry* **281**: 1935-1942.
- Calvo, E., Mans, B.J., Ribeiro, J.M., and Andersen, J.F. 2009. Multifunctionality and mechanism of ligand binding in a mosquito antiinflammatory protein. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 3728-3733.
- Campanacci, V., Spinelli, S., Lartigue, A., Lewandowski, C., Brown, K., Tegoni, M., and Cambillau, C. 2001. Recombinant chemosensory protein (CSP2) from the moth *Mamestra brassicae*: crystallization and preliminary crystallographic study. *Acta crystallographica* **57**: 137-139.
- Carey, A.F., Wang, G., Su, C.Y., Zwiebel, L.J., and Carlson, J.R. 2010. Odorant reception in the malaria mosquito *Anopheles gambiae*. *Nature* **464**: 66-71.

-
- Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. 2005. The Amber biomolecular simulation programs. *Journal of computational chemistry* **26**: 1668-1688.
- Charifson, P.S., Corkery, J.J., Murcko, M.A., and Walters, W.P. 1999. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of medicinal chemistry* **42**: 5100-5109.
- Chon JK, C.J., Kim SS, Shin W. 2005. Classification of peroxiredoxin subfamilies using regular expressions. *Genomics informatics* **3**: 55-60.
- Chothia, C. and Lesk, A.M. 1986. The relation between the divergence of sequence and structure in proteins. *The EMBO journal* **5**: 823-826.
- Choumet, V., Carmi-Leroy, A., Laurent, C., Lenormand, P., Rousselle, J.C., Namane, A., Roth, C., and Brey, P.T. 2007. The salivary glands and saliva of *Anopheles gambiae* as an essential step in the Plasmodium life cycle: a global proteomic study. *Proteomics* **7**: 3384-3394.
- Collot-Teixeira, S., Martin, J., McDermott-Roe, C., Poston, R., and McGregor, J.L. 2007. CD36 and macrophages in atherosclerosis. *Cardiovascular research* **75**: 468-477.
- Cork, A. 1996. Olfactory basis of host location by mosquitoes and other haematophagous Diptera. *Ciba Foundation symposium* **200**: 71-84; discussion 84-78.
- Cork, A. and Park, K.C. 1996. Identification of electrophysiologically-active compounds for the malaria mosquito, *Anopheles gambiae*, in human sweat extracts. *Medical and veterinary entomology* **10**: 269-276.
- Costantini, C., Birkett, M.A., Gibson, G., Ziesmann, J., Sagnon, N.F., Mohammed, H.A., Coluzzi, M., and Pickett, J.A. 2001. Electroantennogram and behavioural responses of the malaria vector *Anopheles gambiae* to human-specific sweat components. *Medical and veterinary entomology* **15**: 259-266.
- Czarnowski, D. and Gorski, J. 1991. Sweat ammonia excretion during submaximal cycling exercise. *J Appl Physiol* **70**: 371-374.
- Damberger, F., Nikonova, L., Horst, R., Peng, G., Leal, W.S., and Wuthrich, K. 2000. NMR characterization of a pH-dependent equilibrium between two folded solution conformations of the pheromone-binding protein from *Bombyx mori*. *Protein Sci* **9**: 1038-1041.
- Damberger, F.F., Ishida, Y., Leal, W.S., and Wuthrich, K. 2007. Structural basis of ligand binding and release in insect pheromone-binding proteins: NMR structure of *Antheraea polyphemus* PBP1 at pH 4.5. *Journal of molecular biology* **373**: 811-819.
- Davrazou, F., Dong, E., Murphy, E.J., Jhonsom, H.T., and Jones, D.N.M. 2011. New insights into the mechanism of odorant detection by the malaria transmitting mosquito *Anopheles gambiae*. *Journal of Biological Chemistry*.
- Dayhoff, M.O., Barker, W.C., and Hunt, L.T. 1983. Establishing homologies in protein sequences. *Methods in enzymology* **91**: 524-545.
- Dietz, V., Gubler, D.J., Ortiz, S., Kuno, G., Casta-Velez, A., Sather, G.E., Gomez, I., and Vergne, E. 1996. The 1986 dengue and dengue hemorrhagic fever epidemic in Puerto Rico: epidemiologic and clinical observations. *Puerto Rico health sciences journal* **15**: 201-210.
- Ellin, R.I., Farrand, R.L., Oberst, F.W., Crouse, C.L., Billups, N.B., Koon, W.S., Musselman, N.P., and Sidell, F.R. 1974. An apparatus for the detection and quantitation of volatile human effluents. *J Chromatogr* **100**: 137-152.
- Fabre, J. 1916. *The life of the Caterpillar*. New York: Dodd, Mead and Company.
- Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**.
-

- Fetrow, J.S. and Skolnick, J. 1998. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *Journal of molecular biology* **281**: 949-968.
- Fiser, A., Do, R.K., and Sali, A. 2000. Modeling of loops in protein structures. *Protein Sci* **9**: 1753-1773.
- Fiser, A., Feig, M., Brooks, C.L., 3rd, and Sali, A. 2002. Evolution and physics in comparative protein structure modeling. *Accounts of chemical research* **35**: 413-421.
- Foster, W.A. 1995. Mosquito sugar feeding and reproductive energetics. *Annual review of entomology* **40**: 443-474.
- Fradin, M.S. and Day, J.F. 2002. Comparative efficacy of insect repellents against mosquito bites. *The New England journal of medicine* **347**: 13-18.
- Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelley, M., Perry, J.K. et al. 2004. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry* **47**: 1739-1749.
- Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., Greenwood, J.R., Halgren, T.A., Sanschagrin, P.C., and Mainz, D.T. 2006. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of medicinal chemistry* **49**: 6177-6196.
- Galindo, K. and Smith, D. 2001. A large family of divergent *Drosophila* odorant-binding proteins expressed in gustatory and olfactory sensilla. *Genetics* **159**: 1059 - 1072.
- Gilles, M.T. 1954. The adult stages of *Prospistoma* Latreille (Emphemeroptera), with descriptions of two new species from africa. *Transactions of the royal entomological society of London* **105**: 355-372.
- Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L. 2003. 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics (Oxford, England)* **19**: 1015-1018.
- Githeko, A.K., Lindsay, S.W., Confalonieri, U.E., and Patz, J.A. 2000. Climate change and vector-borne diseases: a regional analysis. *Bulletin of the World Health Organization* **78**: 1136-1147.
- Godzik, A. 1996. Knowledge-based potentials for protein folding: what can we learn from known protein structures? *Structure* **4**: 363-366.
- Gonnet, G.H., Cohen, M.A., and Benner, S.A. 1992. Exhaustive matching of the entire protein sequence database. *Science (New York, N.Y)* **256**: 1443-1445.
- Graham, L.A. and Davies, P.L. 2002. The odorant-binding proteins of *Drosophila melanogaster*: annotation and characterization of a divergent gene family. *Gene* **292**: 43-55.
- Grater, F., de Groot, B.L., Jiang, H., and Grubmuller, H. 2006. Ligand-release pathways in the pheromone-binding protein of *Bombyx mori*. *Structure* **14**: 1567-1576.
- Gribskov, M. and Veretnik, S. 1996. Identification of sequence pattern with profile analysis. *Methods in enzymology* **266**: 198-212.
- Ha, T.S. and Smith, D.P. 2006. A pheromone receptor mediates 11-cis-vaccenyl acetate-induced responses in *Drosophila*. *J Neurosci* **26**: 8727-8733.
- Hales, S., de Wet, N., Maindonald, J., and Woodward, A. 2002. Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *Lancet* **360**: 830-834.
- Hallem, E.A. and Carlson, J.R. 2006. Coding of odors by a receptor repertoire. *Cell* **125**: 143-160.
- Halperin, I., Ma, B., Wolfson, H., and Nussinov, R. 2002. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* **47**: 409-443.

-
- Harrison, P.M. and Sternberg, M.J. 1994. Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage. *Journal of molecular biology* **244**: 448-463.
- Healy, T.P. and Copland, M.J. 2000. Human sweat and 2-oxopentanoic acid elicit a landing response from *Anopheles gambiae*. *Medical and veterinary entomology* **14**: 195-200.
- Healy, T.P., Copland, M.J., Cork, A., Przyborowska, A., and Halket, J.M. 2002. Landing responses of *Anopheles gambiae* elicited by oxocarboxylic acids. *Medical and veterinary entomology* **16**: 126-132.
- Hekmat-Scafe, D.S., Scafe, C.R., McKinney, A.J., and Tanouye, M.A. 2002. Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. *Genome research* **12**: 1357-1369.
- Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America* **89**: 10915-10919.
- Henikoff, S. and Henikoff, J.G. 1993. Performance evaluation of amino acid substitution matrices. *Proteins* **17**: 49-61.
- Hill, C.A., Fox, A.N., Pitts, R.J., Kent, L.B., Tan, P.L., Chrystal, M.A., Cravchik, A., Collins, F.H., Robertson, H.M., and Zwiebel, L.J. 2002. G protein-coupled receptors in *Anopheles gambiae*. *Science (New York, N.Y)* **298**: 176-178.
- Honson, N., Johnson, M.A., Oliver, J.E., Prestwich, G.D., and Plettner, E. 2003. Structure-activity studies with pheromone-binding proteins of the gypsy moth, *Lymantria dispar*. *Chemical senses* **28**: 479-489.
- Horst, R., Damberger, F., Luginbuhl, P., Guntert, P., Peng, G., Nikonova, L., Leal, W.S., and Wuthrich, K. 2001. NMR structure reveals intramolecular regulation mechanism for pheromone binding and release. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 14374-14379.
- Hubbard, T.J. and Blundell, T.L. 1987. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein engineering* **1**: 159-171.
- Humphrey, W., Dalke, A., and Schulten, K. 1996. VMD: visual molecular dynamics. *Journal of molecular graphics* **14**: 33-38, 27-38.
- Innis, C.A., Anand, A.P., and Sowdhamini, R. 2004. Prediction of functional sites in proteins using conserved functional group analysis. *Journal of molecular biology* **337**: 1053-1068.
- Jaroszewski, L., Rychlewski, L., Zhang, B., and Godzik, A. 1998. Fold prediction by a hierarchy of sequence, threading, and modeling methods. *Protein Sci* **7**: 1431-1440.
- Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., and Gibson, T.J. 1998. Multiple sequence alignment with Clustal X. *Trends in biochemical sciences* **23**: 403-405.
- Jennings, A.J., Edge, C.M., and Sternberg, M.J. 2001. An approach to improving multiple alignments of protein sequences using predicted secondary structure. *Protein engineering* **14**: 227-231.
- Jin, X., Ha, T.S., and Smith, D.P. 2008. SNMP is a signaling component required for pheromone sensitivity in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 10996-11001.
- Johnson, M.S. and Overington, J.P. 1993. A structural basis for sequence comparisons. An evaluation of scoring methodologies. *Journal of molecular biology* **233**: 716-738.
- Johnson, M.S., Srinivasan, N., Sowdhamini, R., and Blundell, T.L. 1994. Knowledge-based protein modeling. *Critical reviews in biochemistry and molecular biology* **29**: 1-68.
- Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**: 275-282.
-

- Kalume, D.E., Okulate, M., Zhong, J., Reddy, R., Suresh, S., Deshpande, N., Kumar, N., and Pandey, A. 2005. A proteomic analysis of salivary glands of female *Anopheles gambiae* mosquito. *Proteomics* **5**: 3765-3777.
- Keisuke Watanabe, Yoshinori Shono, Akiko Kakimizu, Akihiko Okada, Noritada Matsuo, Atsushi Satoh, and Nishimurag, H. 1993. New Mosquito Repellent from *Eucalyptus camaldulensis*. *J. Agric. Food Chem* **41**: 2164-2166.
- Khodade, P., Prabhu, R., Chandra, N., Raha, S., and Govindarajan, R. 2007. Parallel implementation of Autodock. *Journal of Applied crystallography* **40**: 598-599.
- Kim, D.H., Kim, S.I., Chang, K.S., and Ahn, Y.J. 2002. Repellent activity of constituents identified in *Foeniculum vulgare* fruit against *Aedes aegypti* (Diptera: Culicidae). *Journal of agricultural and food chemistry* **50**: 6993-6996.
- Kim, M.S., Repp, A., and Smith, D.P. 1998. LUSH odorant-binding protein mediates chemosensory responses to alcohols in *Drosophila melanogaster*. *Genetics* **150**: 711-721.
- Kostelc, J.G., Preti, G., Zelson, P.R., Stoller, N.H., and Tonzetich, J. 1980. Salivary volatiles as indicators of periodontitis. *Journal of periodontal research* **15**: 185-192.
- Kraut, J. 1977. Serine proteases: structure and mechanism of catalysis. *Annual review of biochemistry* **46**: 331-358.
- Krieger, J., Ganssle, H., Raming, K., and Breer, H. 1993. Odorant binding proteins of *Heliothis virescens*. *Insect biochemistry and molecular biology* **23**: 449-456.
- Krieger, J., Grosse-Wilde, E., Gohl, T., and Breer, H. 2005. Candidate pheromone receptors of the silkmoth *Bombyx mori*. *The European journal of neuroscience* **21**: 2167-2176.
- Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *Journal of molecular biology* **235**: 1501-1531.
- Kruse, S.W., Zhao, R., Smith, D.P., and Jones, D.N. 2003. Structure of a specific alcohol-binding site defined by the odorant binding protein LUSH from *Drosophila melanogaster*. *Nature structural biology* **10**: 694-700.
- Kumar, S., Tamura, K., and Nei, M. 1994. MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput Appl Biosci* **10**: 189-191.
- Kurtovic, A., Widmer, A., and Dickson, B.J. 2007. A single class of olfactory neurons mediates behavioural responses to a *Drosophila* sex pheromone. *Nature* **446**: 542-546.
- Kwon, H.W., Lu, T., Rutzler, M., and Zwiebel, L.J. 2006. Olfactory responses in a gustatory organ of the malaria vector mosquito *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 13526-13531.
- Lagarde, A., Spinelli, S., Qiao, H., Tegoni, M., Pelosi, P., and Cambillau, C. 2011. Crystal structure of a novel type of odorant-binding protein from *Anopheles gambiae*, belonging to the C-plus class. *The Biochemical journal* **437**: 423-430.
- Lartigue, A., Gruez, A., Briand, L., Blon, F., Bezirard, V., Walsh, M., Pernollet, J.C., Tegoni, M., and Cambillau, C. 2004. Sulfur single-wavelength anomalous diffraction crystal structure of a pheromone-binding protein from the honeybee *Apis mellifera* L. *The Journal of biological chemistry* **279**: 4459-4464.
- Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. 1996. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *Journal of biomolecular NMR* **8**: 477-486.
- Laughlin, J.D., Ha, T.S., Jones, D.N., and Smith, D.P. 2008. Activation of pheromone-sensitive neurons is mediated by conformational activation of pheromone-binding protein. *Cell* **133**: 1255-1265.

- Lautenschlager, C., Leal, W.S., and Clardy, J. 2005. Coil-to-helix transition and ligand release of *Bombyx mori* pheromone-binding protein. *Biochemical and biophysical research communications* **335**: 1044-1050.
- Lautenschlager, C., Leal, W.S., and Clardy, J. 2007. *Bombyx mori* pheromone-binding protein binding nonpheromone ligands: implications for pheromone recognition. *Structure* **15**: 1148-1154.
- Lawson, D., Arensburger, P., Atkinson, P., Besansky, N.J., Bruggner, R.V., Butler, R., Campbell, K.S., Christophides, G.K., Christley, S., Dialynas, E. et al. 2009. VectorBase: a data resource for invertebrate vector genomics. *Nucleic acids research* **37**: D583-587.
- Lee, D., Damberger, F.F., Peng, G., Horst, R., Guntert, P., Nikonova, L., Leal, W.S., and Wuthrich, K. 2002. NMR structure of the unliganded *Bombyx mori* pheromone-binding protein at physiological pH. *FEBS letters* **531**: 314-318.
- Leite, N.R., Krogh, R., Xu, W., Ishida, Y., Iulek, J., Leal, W.S., and Oliva, G. 2009. Structure of an odorant-binding protein from the mosquito *Aedes aegypti* suggests a binding pocket covered by a pH-sensitive "Lid". *PloS one* **4**: e8006.
- Lenffer, J., Lai, P., El Mejaber, W., Khan, A.M., Koh, J.L., Tan, P.T., Seah, S.H., and Brusich, V. 2004. CysView: protein classification based on cysteine pairing patterns. *Nucleic acids research* **32**: W350-355.
- Lesk, A.M. and Chothia, C. 1980. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of molecular biology* **136**: 225-270.
- Li, H., Robertson, A.D., and Jensen, J.H. 2005. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* **61**: 704-721.
- Lindh, J.M., Kannaste, A., Knols, B.G., Faye, I., and Borg-Karlson, A.K. 2008. Oviposition responses of *Anopheles gambiae* s.s. (Diptera: Culicidae) and identification of volatiles from bacteria-containing solutions. *Journal of medical entomology* **45**: 1039-1049.
- Logan, J.G., Birkett, M.A., Clark, S.J., Powers, S., Seal, N.J., Wadhams, L.J., Mordue Luntz, A.J., and Pickett, J.A. 2008. Identification of human-derived volatile chemicals that interfere with attraction of *Aedes aegypti* mosquitoes. *J Chem Ecol* **34**: 308-322.
- Lu, T., Qiu, Y.T., Wang, G., Kwon, J.Y., Rutzler, M., Kwon, H.W., Pitts, R.J., van Loon, J.J., Takken, W., Carlson, J.R. et al. 2007. Odor coding in the maxillary palp of the malaria vector mosquito *Anopheles gambiae*. *Curr Biol* **17**: 1533-1544.
- Malnic, B., Hirono, J., Sato, T., and Buck, L.B. 1999. Combinatorial receptor codes for odors. *Cell* **96**: 713-723.
- Mans, B.J., Ribeiro, J.M., and Andersen, J.F. 2008. Structure, function, and evolution of biogenic amine-binding proteins in soft ticks. *The Journal of biological chemistry* **283**: 18721-18733.
- Mao, Y., Xu, X., Xu, W., Ishida, Y., Leal, W.S., Ames, J.B., and Clardy, J. 2010 . Crystal and solution structures of an odorant-binding protein from the southern house mosquito complexed with an oviposition pheromone. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 19102-19107.
- Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure* **29**: 291-325.
- Mas, J.M., Aloy, P., Martı́n-Renom, M. A., Oliva, B., de Llorens, R., Avilés, and F. X. & Querol, E. 2001. Classification of protein disulphide-bridge topologies. *Journal of Computer-Aided Molecular Design* **15**: 477-487.

- McKenna, M.P., Hekmat-Scafe, D.S., Gaines, P., and Carlson, J.R. 1994. Putative *Drosophila* pheromone-binding proteins expressed in a subregion of the olfactory system. *The Journal of biological chemistry* **269**: 16340-16347.
- Meijerink, J., Braks, M.A.H., Brack, A.A., Adam, W., Dekker, T., Posthumus, M.A., Van Beek, T.A., and an Loon, J.J.A. 2000. Identification of Olfactory Stimulants for *Anopheles gambiae* from Human Sweat Samples. *Journal of Chemical Ecology* **26**: 1367-1382.
- Millar, J.G., Chaney, J.D., and Mulla, M.S. 1992. Identification of oviposition attractants for *Culex quinquefasciatus* from fermented Bermuda grass infusions. *Journal of the American Mosquito Control Association* **8**: 11-17.
- Miller, J., McLachlan, A.D., and Klug, A. 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *The EMBO journal* **4**: 1609-1614.
- Mohanty, S., Zubkov, S., and Gronenborn, A.M. 2004. The solution NMR structure of *Antheraea polyphemus* PBP provides new insight into pheromone recognition by pheromone-binding proteins. *Journal of molecular biology* **337**: 443-451.
- Mohl, C., Breer, H., and Krieger, J. 2002. Species-specific pheromonal compounds induce distinct conformational changes of pheromone binding protein subtypes from *Antheraea polyphemus*. *Invert Neurosci* **4**: 165-174.
- Mondal, S., Bhavna, R., Mohan Babu, R., and Ramakumar, S. 2006. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *Journal of theoretical biology* **243**: 252-260.
- Moreno-Hagelsieb, G. and Latimer, K. 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics (Oxford, England)* **24**: 319-324.
- Morris, A.J. and Zhang, J. 1998. A Sequential Learning Approach for Single Hidden Layer Neural Networks. *Neural Netw* **11**: 65-80.
- Muegge, I., Heald, S.L., and Brittelli, D. 2001. Simple selection criteria for drug-like chemical matter. *Journal of medicinal chemistry* **44**: 1841-1846.
- Needleman, S.B. and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **48**: 443-453.
- Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C., Tu, Z.J., Loftus, B., Xi, Z., Megy, K., Grabherr, M. et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science (New York, N.Y)* **316**: 1718-1723.
- Nissink, J.W. 2009. Simple size-independent measure of ligand efficiency. *Journal of chemical information and modeling* **49**: 1617-1622.
- O'Brien, K.P., Remm, M., and Sonnhammer, E.L. 2005. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research* **33**: D476-480.
- Olivera, B.M. 2002. Conus venom peptides : Reflections from the biology of clades and species. *Annu Rev Ecol. syst* **33**: 25-47.
- Omolo, M.O., Okinyo, D., Ndiege, I.O., Lwande, W., and Hassanali, A. 2004. Repellency of essential oils of some Kenyan plants against *Anopheles gambiae*. *Phytochemistry* **65**: 2797-2802.
- Ouzounis, C.A., Coulson, R.M., Enright, A.J., Kunin, V., and Pereira-Leal, J.B. 2003. Classification schemes for protein structure and function. *Nat Rev Genet* **4**: 508-519.
- Overington, J., Donnelly, D., Johnson, M.S., Sali, A., and Blundell, T.L. 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci* **1**: 216-226.
- Paquet C, Quatresous I, Solet JL, Sissoko D, Renault P, Pierre V, Cordel H, Lassalle C, Thiria J, Zeller H et al. 2006. Chikungunya outbreak in Réunion: epidemiology and surveillance, 2005 to early January 2006. *Euro Surveillance* **11**.

-
- Pearson, W.R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods in molecular biology (Clifton, N.J)* **132**: 185-219.
- Pelletier, J., Guidolin, A., Syed, Z., Cornel, A.J., and Leal, W.S. 2010. Knockdown of a mosquito odorant-binding protein involved in the sensitive detection of oviposition attractants. *J Chem Ecol* **36**: 245-248.
- Pelletier, J. and Leal, W.S. 2009. Genome analysis and expression patterns of odorant-binding proteins from the Southern House mosquito *Culex pipiens quinquefasciatus*. *PloS one* **4**: e6237.
- Pelosi, P. and Maida, R. 1995. Odorant-binding proteins in insects. *Comparative biochemistry and physiology* **111**: 503-514.
- Pesenti, M.E., Spinelli, S., Bezirard, V., Briand, L., Pernollet, J.C., Campanacci, V., Tegoni, M., and Cambillau, C. 2009. Queen bee pheromone binding protein pH-induced domain swapping favors pheromone release. *Journal of molecular biology* **390**: 981-990.
- Pesenti, M.E., Spinelli, S., Bezirard, V., Briand, L., Pernollet, J.C., Tegoni, M., and Cambillau, C. 2008. Structural basis of the honey bee PBP pheromone and pH-induced conformational change. *Journal of molecular biology* **380**: 158-169.
- Pikielny, C.W., Hasan, G., Rouyer, F., and Rosbash, M. 1994. Members of a family of *Drosophila* putative odorant-binding proteins are expressed in different subsets of olfactory hairs. *Neuron* **12**: 35-49.
- Plettner, E., Lazar, J., Prestwich, E.G., and Prestwich, G.D. 2000. Discrimination of pheromone enantiomers by two pheromone binding proteins from the gypsy moth *Lymantria dispar*. *Biochemistry* **39**: 8953-8962.
- Price, M.L.P., Ostrovsky, D., and Jorgensen, W.L. 2001. Gas-phase and liquid state properties of esters, nitriles and nitro compounds with the OPLS-AA force field. *Journal of computational chemistry* **22**: 1340-1352.
- Qiao, H., He, X., Schymura, D., Ban, L., Field, L., Dani, F.R., Michelucci, E., Caputo, B., della Torre, A., Iatrou, K. et al. 2010. Cooperative interactions between odorant-binding proteins of *Anopheles gambiae*. *Cell Mol Life Sci* **68**: 1799-1813.
- Qiu, Y.T., van Loon, J.J., Takken, W., Meijerink, J., and Smid, H.M. 2006. Olfactory Coding in Antennal Neurons of the Malaria Mosquito, *Anopheles gambiae*. *Chemical senses* **31**: 845-863.
- Reiter, P., Fontenille, D., and Paupy, C. 2006. *Aedes albopictus* as an epidemic vector of chikungunya virus: another emerging problem? *The Lancet infectious diseases* **6**: 463-464.
- Richardson, J.S. 1981. The anatomy and taxonomy of protein structure. *Advances in protein chemistry* **34**: 167-339.
- Ronderos, D.S. and Smith, D.P. 2009. Diverse signaling mechanisms mediate volatile odorant detection in *Drosophila*. *Fly* **3**: 290-297.
- Ross, R. 2002. The role of the mosquito in the evolution of the malarial parasite: the recent researches of Surgeon-Major Ronald Ross, I.M.S. 1898. *The Yale journal of biology and medicine* **75**: 103-105.
- Rothmund, S., Liou, Y.C., Davies, P.L., Krause, E., and Sonnichsen, F.D. 1999. A new class of hexahelical insect proteins revealed as putative carriers of small hydrophobic ligands. *Structure* **7**: 1325-1332.
- Rueda, L.M., Patel, K.J., Axtell, R.C., and Stinner, R.E. 1990. Temperature-dependent development and survival rates of *Culex quinquefasciatus* and *Aedes aegypti* (Diptera: Culicidae). *Journal of medical entomology* **27**: 892-898.
- Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* **4**: 406-425.
-

- Sali, A. 1995. Comparative protein modeling by satisfaction of spatial restraints. *Molecular medicine today* **1**: 270-277.
- Sali, A. and Blundell, T.L. 1990. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *Journal of molecular biology* **212**: 403-428.
- Sali, A. and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology* **234**: 779-815.
- Sali, A. and L.Blundell, T. 1990. Definition of general topological equivalence in protein strutures . A procedure involving comprision of properties and relationships through simulated annealing and dynamic programming. *Journal of molecular biology* **212**: 403 - 428.
- Sali, A. and Overington, J.P. 1994. Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci* **3**: 1582-1596.
- Sanchez, R., Pieper, U., Mirkovic, N., de Bakker, P.I., Wittenstein, E., and Sali, A. 2000. MODBASE, a database of annotated comparative protein structure models. *Nucleic acids research* **28**: 250-253.
- Sanchez, R. and Sali, A. 1997a. Advances in comparative protein-structure modelling. *Current opinion in structural biology* **7**: 206-214.
- Sanchez, R. and Sali, A. 1997b. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins Suppl* **1**: 50-58.
- Sanchez-Gracia, A., Vieira, F.G., and Rozas, J. 2009. Molecular evolution of the major chemosensory gene families in insects. *Heredity* **103**: 208-216.
- Sandler, B.H., Nikonova, L., Leal, W.S., and Clardy, J. 2000. Sexual attraction in the silkworm moth: structure of the pheromone-binding-protein-bombykol complex. *Chemistry & biology* **7**: 143-151.
- Shi, J., Blundell, T.L., and Mizuguchi, K. 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *Journal of molecular biology* **310**: 243-257.
- Simon, I., Fiser, A., and Tusnady, G.E. 2001. Predicting protein conformation by statistical methods. *Biochimica et biophysica acta* **1549**: 123-136.
- Sippl, M.J. 1995. Knowledge-based potentials for proteins. *Current opinion in structural biology* **5**: 229-235.
- Smallegange, R.C., Qiu, Y.T., van Loon, J.J., and Takken, W. 2005. Synergism between ammonia, lactic acid and carboxylic acids as kairomones in the host-seeking behaviour of the malaria mosquito *Anopheles gambiae sensu stricto* (Diptera: Culicidae). *Chemical senses* **30**: 145-152.
- Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *Journal of molecular biology* **147**: 195-197.
- Srinivasan, N., Sowdhamini, R., Ramakrishnan, C., and Balaram, P. 1990. Conformations of disulfide bridges in proteins. *International Journal of Peptide and Protein Research* **36**: 1399-3011.
- Steinbrecht, R.A. 1998. Odorant-binding proteins: expression and function. *Annals of the New York Academy of Sciences* **855**: 323-332.
- Sutcliffe, M.J., Haneef, I., Carney, D., and Blundell, T.L. 1987. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein engineering* **1**: 377-384.
- Sutcliffe, M.J., Hayes, F.R., and Blundell, T.L. 1987. Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. *Protein engineering* **1**: 385-392.

-
- Swarup, S., Williams, T.I., and Anholt, R.R. 2011. Functional dissection of Odorant binding protein genes in *Drosophila melanogaster*. *Genes, brain, and behavior* **10**: 648-657.
- Takken, W., Dekker, T., and Wijnholds, Y.G. 1997. Odor-mediated flight behavior of *Anopheles gambiae* Giles sensu stricto and *An. stephensi* Liston in response to CO₂, acetone, and 1-octen-3-ol (Diptera: Culicidae). *Journal of insect behavior* **10**.
- Takken, W. and Knols, B.G.J. 1999. Odor mediated behavior of afrotropical malaria mosquitoes. *Annual review of entomology* **44**: 131-157.
- Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular biology and evolution* **24**: 1596-1599.
- Tangerman, A., Meuwese-Arends, M.T., and van Tongeren, J.H. 1983. A new sensitive assay for measuring volatile sulphur compounds in human breath by Tenax trapping and gas chromatography and its application in liver cirrhosis. *Clinica chimica acta; international journal of clinical chemistry* **130**: 103-110.
- Teague, S.J. 2003. Implications of protein flexibility for drug discovery. *Nature reviews* **2**: 527-541.
- Tegoni, M., Campanacci, V., and Cambillau, C. 2004. Structural aspects of sexual attraction and chemical communication in insects. *Trends in biochemical sciences* **29**: 257-264.
- Thangudu, R., Manoharan, M., Srinivasan, N., Cadet, F., SOWDHAMINI, R., and Offmann, B. 2008. Analysis on conservation of disulphide bonds and their structural features in homologous protein domain families. *BMC Structural Biology* **8**.
- Thangudu, R.R., Vinayagam, A., Pugalenthi, G., Manonmani, A., Offmann, B., and Sowdhamini, R. 2005. Native and modeled disulfide bonds in proteins: knowledge-based approaches toward structure prediction of disulfide-rich polypeptides. *Proteins* **58**: 866-879.
- Thode, A.B., Kruse, S.W., Nix, J.C., and Jones, D.N. 2008. The role of multiple hydrogen-bonding groups in specific alcohol binding sites in proteins: insights from structural studies of LUSH. *Journal of molecular biology* **376**: 1360-1376.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic acids research* **25**: 4876-4882.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**: 4673-4680.
- Thornton, J.M. 1981. Disulphide bridges in globular proteins. *Journal of molecular biology* **151**: 261-287.
- Townson, H. and Nathan, M.B. 2008. Resurgence of chikungunya. *Transactions of the royal society of tropical medicine and hygiene* **102**: 308-309.
- Tsitsanou, K.E., Thireou, T., Drakou, C.E., Koussis, K., Keramioti, M.V., Leonidas, D.D., Eliopoulos, E., Iatrou, K., and Zographos, S.E. 2011. *Anopheles gambiae* odorant binding protein crystal complex with the synthetic repellent DEET: implications for structure-based design of novel mosquito repellents. *Cell Mol Life Sci*.
- Turell, M.J. 1989. Effect of environmental temperature on the vector competence of *Aedes fowleri* for Rift Valley fever virus. *Research in virology* **140**: 147-154.
- Turner, S.L. and Ray, A. 2009. Modification of CO₂ avoidance behaviour in *Drosophila* by inhibitory odorants. *Nature* **461**: 277-281.
- Valenzuela, J.G., Pham, V.M., Garfield, M.K., Francischetti, I.M., and Ribeiro, J.M. 2002. Toward a description of the sialome of the adult female mosquito *Aedes aegypti*. *Insect biochemistry and molecular biology* **32**: 1101-1122.
- Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J. 2005. GROMACS: fast, flexible, and free. *Journal of computational chemistry* **26**: 1701-1718.
-

- Vartak, P.H. and Sharma, R.N. 1993. Vapour toxicity & repellence of some essential oils & terpenoids to adults of *Aedes aegypti* (L) (Diptera: Culicidae). *The Indian journal of medical research* **97**: 122-127.
- Verhulst, N.O., Beijleveld, H., Knols, B.G., Takken, W., Schraa, G., Bouwmeester, H.J., and Smallegange, R.C. 2009. Cultured skin microbiota attracts malaria mosquitoes. *Malaria journal* **8**: 302.
- Vieira, F.G. and Rozas, J. 2011. Comparative Genomics of the Odorant-Binding and Chemosensory Protein Gene Families across the Arthropoda: Origin and Evolutionary History of the Chemosensory System. *Genome biology and evolution* **3**: 476-490.
- Vieira, F.G., Sanchez-Gracia, A., and Rozas, J. 2007. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. *Genome biology* **8**: R235.
- Vogt, R.G., Callahan, F.E., Rogers, M.E., and Dickens, J.C. 1999. Odorant binding protein diversity and distribution among the insect orders, as indicated by LAP, an OBP-related protein of the true bug *Lygus lineolaris* (Hemiptera, Heteroptera). *Chemical senses* **24**: 481-495.
- Vogt, R.G. and Riddiford, L.M. 1981. Pheromone binding and inactivation by moth antennae. *Nature* **293**: 161-163.
- Vogt, R.G., Rogers, M.E., Franco, M.D., and Sun, M. 2002. A comparative study of odorant binding protein genes: differential expression of the PBP1-GOBP2 gene cluster in *Manduca sexta* (Lepidoptera) and the organization of OBP genes in *Drosophila melanogaster* (Diptera). *The Journal of experimental biology* **205**: 719-744.
- Wahl, H.G., Hoffmann, A., Luft, D., and Liebich, H.M. 1999. Analysis of volatile organic compounds in human urine by headspace gas chromatography-mass spectrometry with a multipurpose sampler. *Journal of chromatography* **847**: 117-125.
- Walter R. P. Scott, P.H.H., Ilario G. Tironi, Alan E. Mark, Salomon R. Billeter, Jens Fennen, Andrew E. Torda, Thomas Huber, Peter Krüger, and Wilfred F. van Gunsteren. 1999. The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **103**: 3596-3607.
- Wang, G., Carey, A.F., Carlson, J.R., and Zwiebel, L.J. 2010. Molecular basis of odor coding in the malaria vector mosquito *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 4418-4423.
- Wang, Q., Hasan, G., and Pikielny, C.W. 1999. Preferential expression of biotransformation enzymes in the olfactory organs of *Drosophila melanogaster*, the antennae. *The Journal of biological chemistry* **274**: 10309-10315.
- Watson, R.T., M.C.Zinyowera, and R.H.MOSS, . 1996. Climate change 1995: Impacts, Adaptations and Mitigation of Climate change. In *IPCC Working Group I Report*, pp. 55-57. Cambridge University Press.
- William L. Jorgensen, D.S.M., and Julian Tirado-Rives. 1996. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the american chemical society* **118**: 11225-11236.
- Wogulis, M., Morgan, T., Ishida, Y., Leal, W.S., and Wilson, D.K. 2006. The crystal structure of an odorant binding protein from *Anopheles gambiae*: evidence for a common ligand release mechanism. *Biochemical and biophysical research communications* **339**: 157-164.
- Wojtasek, H. and Leal, W.S. 1999. Conformational change in the pheromone-binding protein from *Bombyx mori* induced by pH and by interaction with membranes. *The Journal of biological chemistry* **274**: 30950-30956.
- Xu, P., Atkinson, R., Jones, D.N., and Smith, D.P. 2005. *Drosophila* OBP LUSH is required for activity of pheromone-sensitive neurons. *Neuron* **45**: 193-200.

-
- Xu, P.X., Zwiebel, L.J., and Smith, D.P. 2003. Identification of a distinct family of genes encoding atypical odorant-binding proteins in the malaria vector mosquito, *Anopheles gambiae*. *Insect molecular biology* **12**: 549-560.
- Zeng XN, Leyden JJ, Lawley HJ, Sawano K, Nohara I, and G., P. 1991. Analysis of characteristic odors from human male axillae. . *J Chem Ecol.* **17**: 1469–1492.
- Zhou, J.J., He, X.L., Pickett, J.A., and Field, L.M. 2008. Identification of odorant-binding proteins of the yellow fever mosquito *Aedes aegypti*: genome annotation and comparative analyses. *Insect molecular biology* **17**: 147-163.
- Zhou, J.J., Huang, W., Zhang, G.A., Pickett, J.A., and Field, L.M. 2004. "Plus-C" odorant-binding protein genes in two *Drosophila* species and the malaria mosquito *Anopheles gambiae*. *Gene* **327**: 117-129.
- Zhou, J.J., Vieira, F.G., He, X.L., Smadja, C., Liu, R., Rozas, J., and Field, L.M. 2009. Genome annotation and comparative analyses of the odorant-binding proteins and chemosensory proteins in the pea aphid *Acyrtosiphon pisum*. *Insect molecular biology* **19 Suppl 2**: 113-122.
- Zubkov, S., Gronenborn, A.M., Byeon, I.J., and Mohanty, S. 2005. Structural consequences of the pH-induced conformational switch in *A.polyphemus* pheromone-binding protein: mechanisms of ligand release. *Journal of molecular biology* **354**: 1081-1090.
- Zuckerkindl E and L, P. 1965. Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*: 97-166.

SUPPLEMENTARY MATERIAL

Supplementary Table 1a. Complete list of OBP genes from *Anopheles gambiae* reported in this study. Alongside their identification and chromosomal locations, shown are their phylogenetic clusters and orthologues in *Aedes aegypti* and *Culex quinquefasciatus*. The four newly discovered genes by us are AgamOBP65-68 and are denoted by an (*). The gene names used here do not completely coincide with the gene names in the recently published work by Vieira & Rozas (2011). As an *addendum* to this Table, we are providing in supplementary Table 1b, a comparative analysis of the names between their study and ours. The last four genes were identified by Vieira & Rozas (2011) and were added to this Table at the last moment after renaming them AgamOBP69-72. †Two genes AgamOBP34 and AgamOBP37 share 100% sequence identity but are localized on distinct chromosome segments. We cannot resolve which of these two genes is the true orthologue of AaegOBP106. Two way (1:1) orthologues are indicated by a # sign.

n°	ID	Name	Length	Chromosome	Start	End	Cluster	Orthologue in <i>Ae. aegypti</i>	Orthologue in <i>Culex</i>
1	AGAP003309	AgamOBP1	144	2R	35643035	35644609	OS-E/OS-F	AaegOBP1	CquiOBP1
2	AGAP003306	AgamOBP2	157	2R	35434051	35434604	OS-E/OS-F	AaegOBP60	CquiOBP3
3	AGAP001409	AgamOBP3	153	2R	4210577	4212097	OS-E/OS-F	AaegOBP38	CquiOBP2
4	AGAP010489	AgamOBP4	150	3L	4997901	4998953	LUSH	AaegOBP39	CquiOBP6
5	AGAP009629	AgamOBP5	156	3R	37321107	37322204	LUSH		
6	AGAP003530	AgamOBP6	155	2R	39200905	39201673	OBP19a	AaegOBP27	CquiOBP13
7	AGAP001556	AgamOBP7	154	2R	6152013	6154470	Pbprp1	AaegOBP2	CquiOBP7
8	AGAP000279	AgamOBP8	176	X	5036744	5037431	NOGROUP		
9	AGAP000278	AgamOBP9	139	X	5035248	5036100	Obp99a	AaegOBP22	CquiOBP43
10	AGAP001189	AgamOBP10	131	2R	1034139	1169444	mclassic6	AaegOBP10	CquiOBP24
11	AGAP002025	AgamOBP11	192	2R	14069095	14069749	mclassic9		
12	AGAP002188	AgamOBP12	159	2R	17328044	17328912	NOGROUP		
13	AGAP002905	AgamOBP13	149	2R	29134157	29134861	Pbprp2/5	AaegOBP57	CquiOBP28
13	AGAP002189	AgamOBP14#	188	2R	17331871	17332553	mclassic9	AaegOBP18#	CquiOBP63#
15	AGAP003307	AgamOBP15	147	2R	35436449	35436969	OS-E/OS-F	AaegOBP36	CquiOBP5
16		AgamOBP16		Previously reported but same gene as AgamOBP15, hence not included in this analysis					
17		AgamOBP17		Previously reported but same gene as AgamOBP1, hence not included in this analysis					
18		AgamOBP18		Previously reported but same gene as AgamOBP6, hence not included in this analysis					
19	AGAP004433	AgamOBP19	137	2R	55987079	55987846	OBP19a	AaegOBP56	CquiOBP12
20	AGAP005208	AgamOBP20	142	2L	12288238	12289440	OBP19a	AaegOBP55	CquiOBP11
21	AGAP008398	AgamOBP21	131	3R	10317255	10317835	mclassic5	AaegOBP8	CquiOBP23
22	AGAP010409	AgamOBP22	144	3L	2853087	2853645	mclassic8	AaegOBP81	CquiOBP44
23	AGAP012318	AgamOBP23	131	3L	40168852	40169329	mclassic3	AaegOBP9	CquiOBP22
24	AGAP012319	AgamOBP24	176	3L	40171315	40172237	mclassic2	AaegOBP77	CquiOBP21
25	AGAP012320	AgamOBP25	142	3L	40209434	40210326	mclassic3	AaegOBP11	CquiOBP19
26	AGAP012321	AgamOBP26	131	3L	40213816	40214477	mclassic3	AaegOBP35	CquiOBP20
27	AGAP012323	AgamOBP27	134	3L	40218226	40218764	mclassic1	AaegOBP65	CquiOBP16

Supplementary Table 2.1a

n°	ID	Name	Length	Chromosome	Start	End	Cluster	Orthologue in <i>Ae. aegyptii</i>	Orthologue in <i>Culex</i>
28	AGAP012325	AgamOBP28	134	3L	40221011	40221620	mclassic3	AaegOBP12	CquiOBP18
29	AGAP012331	AgamOBP29	176	3L	40708383	40709402	Obp59a	AaegOBP83	CquiOBP55
30	AGAP011647	AgamOBP30	289	3L	30753689	30754558	matype4		
31	AGAP010649	AgamOBP31	313	3L	7968501	7969835	matype3	AaegOBP29	
32	AGAP000638	AgamOBP32	320	X	11413679	11414777	matype4	AaegOBP102	
33	AGAP000640	AgamOBP33	334	X	11416806	11417810	NOGROUP		
34	AGAP000644	AgamOBP34†	311	X	11428745	11429680	matype2	AaegOBP106	
35	AGAP000642	AgamOBP35	275	X	11422726	11423559	matype2		
36	AGAP000643	AgamOBP36	275	X	11426182	11427015	matype2		
37	AGAP000641	AgamOBP37†	311	X	11419447	11420382	matype2	AaegOBP106	
38	AGAP000580	AgamOBP38	336	X	10205844	10206981	NOGROUP		
39	AGAP002190	AgamOBP39	246	2R	17333035	17333892	matype1	AaegOBP40	CquiOBP84
40	AGAP002191	AgamOBP40	282	2R	17334179	17335132	matype1		
41	AGAP005182	AgamOBP41	279	2L	11685185	11686024	NOGROUP		
42	AGAP009065	AgamOBP42	288	3R	25310798	25311664	NOGROUP		
43	AGAP009402	AgamOBP43	333	3R	32225167	32226168	matype4		CquiOBP76
44	AGAP010648	AgamOBP44	327	3L	7964187	7965356	matype3		
45	AGAP010650	AgamOBP45	356	3L	7970287	7971357	matype3		
46	AGAP007289	AgamOBP46	202	2L	45014419	45015087	mplus2	AaegOBP48	CquiOBP108
47	AGAP007287	AgamOBP47	228	2L	45011203	45012023	mplus1		
48	AGAP007286	AgamOBP48	200	2L	45008928	45009814	mplus1	AaegOBP42	CquiOBP106
49	AGAP006075	AgamOBP49	179	2L	26103215	26103894	mplus7	AaegOBP73	CquiOBP112
50	AGAP006076	AgamOBP50	166	2L	26104148	26104931	mplus7		
51	AGAP006077	AgamOBP51	203	2L	26129377	26130091	mplus7		
52	AGAP006078	AgamOBP52	170	2L	26130285	26131425	mplus7		
53	AGAP006079	AgamOBP53	171	2L	26131618	26132453	mplus7		
54	AGAP006080	AgamOBP54#	181	2L	26133006	26133968	mplus7	AaegOBP69#	CquiOBP111#
55	AGAP006081	AgamOBP55	156	2L	26134275	26135119	mplus8		
56	AGAP011367	AgamOBP56	235	3L	22025830	22026752	mplus9	AaegOBP26	CquiOBP102
57	AGAP011368	AgamOBP57	204	3L	22028532	22029250	mplus9	AaegOBP25	CquiOBP104
58	AGAP006074	AgamObp59a	286	2L	26101764	26135649	mplus6	AaegOBP5	CquiOBP109
59	AGAP006760	AgamOBP59	155	2L	37923144	37923831	NOGROUP		
60	AGAP007281	AgamOBP60#	198	2L	44998234	44999054	mplus4	AaegOBP51#	CquiOBP110#
61	AGAP007282	AgamOBP61	204	2L	44999450	45000275	mplus4		
62	AGAP002556	AgamOBP62	174	2R	22787889	22792798	Pbprp4	AaegOBP20	CquiOBP14
63	AGAP012322	AgamOBP63	135	3L	40217381	40217853	mclassic4	AaegOBP61	CquiOBP17

n°	ID	Name	Length	Chromosome	Start	End	Cluster	Orthologue in <i>Ae. aegyptii</i>	Orthologue in <i>Culex</i>
64	AGAP012324	AgamOBP64	142	3L	40219631	40220284	mclassic1	AaegOBP15	CquiOBP15
65	AGAP006759	AgamOBP65*	179	2L	37922005	37922885	NOGROUP		
66	AGAP007283	AgamOBP66*	212	2L	45000529	45001413	mplus8		
67	AGAP012659	AgamOBP67*	204	UNKN	22699010	22699835	mplus4		
68	AGAP012658	AgamOBP68*	212	UNKN	22697816	22698745	mplus8		
69	AGAP013182	AgamOBP69	229	2R	21343359	21344048	nd		
70	AGAP006368	AgamOBP70	200	2L	30543243	30547869	nd		CquiOBP114
71	AGAP012867	AgamOBP71	228	UNKN	35107623	35108585	nd		
72	AGAP012714	AgamOBP72	121	UNKN	24728945	24729493	nd		

*Newly discovered genes in this study.

Only a two way (1:1) orthology has been established for these genes and not a three way (1:1:1) orthology.

†Two genes AgamOBP34 and AgamOBP37 share 100% sequence identity but are localized on distinct chromosome segments. We cannot resolve which of these two genes is the true orthologue of AaegOBP106.

nd : not determined

Table 1b. OBP gene names for *Anopheles gambiae*. Shown is the correspondence between the OBP gene names used in this study and the names used in the Vectorbase, Xu et al. 2003 and very recent work of Vieira & Rozas (2011).

n°	Gene identifier in Xu et al. (2003)	Corresponding vectorbase ID	Name in vectorbase	Name as defined in Xu et al. 2003	Name in (Vieira & Rozas. 2011)	Name in (Manoharan et al. 2011)	Comment
1	AY146721	AGAP003309	AgamOBP17	AgamOBP1	AgamOBP1	AgamOBP1	Gene name difference in vectorbase
2	AY146719	AGAP003306	AgamOBP2	AgamOBP2	AgamOBP2	AgamOBP2	
3	AY146745	AGAP001409	AgamOBP3	AgamOBP3	AgamOBP3	AgamOBP3	
4	AY146731	AGAP010489	AgamOBP4	AgamOBP4	AgamOBP4	AgamOBP4	
5	AY146729	AGAP009629	AgamOBP5	AgamOBP5	AgamOBP5	AgamOBP5	
6	AY146725	AGAP003530	AgamOBP6	AgamOBP6	AgamOBP6	AgamOBP6	
7	AY146742	AGAP001556	AgamOBP7	AgamOBP7	AgamOBP7	AgamOBP7	
8	AY146744	AGAP000279	AgamOBP8	AgamOBP8	AgamOBP8	AgamOBP8	
9	AY146740	AGAP000278	AgamOBP9	AgamOBP9	AgamOBP9	AgamOBP9	
10	AY146741	AGAP001189#	AgamOBP10	AgamOBP10	AgamOBP10	AgamOBP10	
11	AY146743	AGAP002025	AgamOBP11	AgamOBP11	AgamOBP11	AgamOBP11	
12	AY146716	AGAP002188	AgamOBP12	AgamOBP12	AgamOBP12	AgamOBP12	
13	AY146718	AGAP002905	AgamOBP14	AgamOBP13	AgamOBP13	AgamOBP13	Gene name difference in vectorbase
14	AY146717	AGAP002189	AgamOBP13	AgamOBP14	AgamOBP14	AgamOBP14	Gene name difference in vectorbase
15	AY146720	AGAP003307	AgamOBP15	AgamOBP15	AgamOBP15	AgamOBP15	
16	AY146722	Not available	Not available	AgamOBP16	AgamOBP16		Not included in this study
17	AY146723	Not available	Not available	AgamOBP17	AgamOBP17		Not included in this study
18	AY146724	Not available	Not available	AgamOBP18	AgamOBP18		Not included in this study
19	AY146726	AGAP004433	AgamOBP19	AgamOBP19	AgamOBP19	AgamOBP19	
20	AY146727	AGAP005208	AgamOBP20	AgamOBP20	AgamOBP20	AgamOBP20	
21	AY146728	AGAP008398	AgamOBP21	AgamOBP21	AgamOBP21	AgamOBP21	
22	AY146730	AGAP010409	AgamOBP22	AgamOBP22	AgamOBP22	AgamOBP22	
23	AY146733	AGAP012318	AgamOBP23	AgamOBP23	AgamOBP23	AgamOBP23	
24	AY146734	AGAP012319	AgamOBP18	AgamOBP24	AgamOBP24	AgamOBP24	Gene name difference in vectorbase
25	AY146735	AGAP012320	AgamOBP25	AgamOBP25	AgamOBP25	AgamOBP25	
26	AY146736	AGAP012321	AgamOBP26	AgamOBP26	AgamOBP26	AgamOBP26	
27	AY146737	AGAP012323	AgamOBP27	AgamOBP27	AgamOBP27	AgamOBP27	
28	AY146738	AGAP012325	AgamOBP28	AgamOBP28	AgamOBP28	AgamOBP28	
29	AY146739	AGAP012331	No name	AgamOBP29	AgamOBP29	AgamOBP29	
30	AY146758	AGAP011647	No name	AgamOBP30	AgamOBP30	AgamOBP30	
31	AY146760	AGAP010649	No name	AgamOBP31	AgamOBP31	AgamOBP31	

Supplementary Table 2.1b

n°	Gene identifier in Xu et al. (2003)	Corresponding vectorbase ID	Name in vectorbase	Name as defined in Xu et al. 2003	Name in (Vieira & Rozas. 2011)	Name in (Manoharan et al. 2011)	Comment
32	AY146755	AGAP000638	AgamOBP32	AgamOBP32	AgamOBP32	AgamOBP32	
33	AY146754	AGAP000640	AgamOBP33	AgamOBP33	AgamOBP33	AgamOBP33	
34	AY146753	AGAP000641	AgamOBP34	AgamOBP34	AgamOBP34	AgamOBP34	
35	AY146752	AGAP000642	AgamOBP35	AgamOBP35	AgamOBP35	AgamOBP35	
36	AY146751	AGAP000643	AgamOBP36	AgamOBP36	AgamOBP36	AgamOBP36	
37	AY146750	AGAP000644	No name	AgamOBP37	AgamOBP37	AgamOBP37	
38	AY146749	AGAP000580	AgamOBP38	AgamOBP38	AgamOBP38	AgamOBP38	
39	AY146757	AGAP002190	AgamOBP39	AgamOBP39	AgamOBP39	AgamOBP39	
40	AY146756	AGAP002191#	AgamOBP40	AgamOBP40	AgamOBP40	AgamOBP40	
41	AY146748	AGAP005182	AgamOBP41	AgamOBP41	AgamOBP41	AgamOBP41	
42	AY146747	AGAP009065	AgamOBP42	AgamOBP42	AgamOBP42	AgamOBP42	
43	AY146746	AGAP009402	AgamOBP43	AgamOBP43	AgamOBP43	AgamOBP43	
44	AY146732	AGAP010648	AgamOBP44	AgamOBP44	AgamOBP44	AgamOBP44	
45	AY146759	AGAP010650	AgamOBP45	AgamOBP45	AgamOBP45	AgamOBP45	
46	AY330173	AGAP007289	AgamOBP46	AgamOBP46	AgamOBP46	AgamOBP46	
47	AY330174	AGAP007287	AgamOBP47	AgamOBP47	AgamOBP47	AgamOBP47	
48	AY330175	AGAP007286	No name	AgamOBP48	AgamOBP48	AgamOBP48	
49	AY330176	AGAP006075	AgamOBP49	AgamOBP49	AgamOBP49	AgamOBP49	
50	AY330177	AGAP006076	AgamOBP50	AgamOBP50	AgamOBP50	AgamOBP50	
51	AY330178	AGAP006077	AgamOBP51	AgamOBP51	AgamOBP51	AgamOBP51	
52	AY330172	AGAP006078	AgamOBP52	AgamOBP52	AgamOBP52	AgamOBP52	
53	AY330179	AGAP006079	AgamOBP53	AgamOBP53	AgamOBP53	AgamOBP53	
54	AY330180	AGAP006080	AgamOBP54	AgamOBP54	AgamOBP54	AgamOBP54	
55	AY330181	AGAP006081	AgamOBP55	AgamOBP55	AgamOBP55	AgamOBP55	
56	AY330182	AGAP011367	AgamOBP56	AgamOBP56	AgamOBP56	AgamOBP56	
57	AY330183	AGAP011368	AgamOBP57	AgamOBP57	AgamOBP57	AgamOBP57	
58		AGAP006074#	Not available	Not available	AgamOBP77	AgamOBP58	Gene name difference in between Vieira & Rozas (2011) and Manoharan et al (2011)
59		AGAP006760#	Not available	Not available	AgamOBP63	AgamOBP59	
60		AGAP007281#	Not available	Not available	AgamOBP64	AgamOBP60	
61		AGAP007282#	Not available	Not available	AgamOBP61	AgamOBP61	
62		AGAP002556#	Not available	Not available	AgamOBP60	AgamOBP62	Gene name difference in between Vieira & Rozas (2011) and Manoharan et al (2011)
63		AGAP012322#	Not available	Not available	AgamOBP58	AgamOBP63	
64		AGAP012324#	Not available	Not available	AgamOBP59	AgamOBP64	

Supplementary Table 2.1b

n°	Gene identifier in Xu et al. (2003)	Corresponding vectorbase ID	Name in vectorbase	Name as defined in Xu et al. 2003	Name in (Vieira & Rozas. 2011)	Name in (Manoharan et al. 2011)	Comment
65		AGAP006759#	Not available	Not available	AgamOBP66	AgamOBP65	Gene name difference in between Vieira & Rozas (2011) and Manoharan et al (2011)
66		AGAP007283#	Not available	Not available	AgamOBP62	AgamOBP66	
67		AGAP012659#	Not available	Not available	AgamOBP67	AgamOBP67	
68		AGAP012658#	Not available	Not available	AgamOBP68	AgamOBP68	
69		AGAP013182#	Not available	Not available	AgamOBP79	AgamOBP69†	New gene identified by Vieira & Rozas (2011). We suggest a new name following the naming in this study
70		AGAP006368#	Not available	Not available	AgamOBP80	AgamOBP70†	
71		AGAP012867#	Not available	Not available	AgamOBP82	AgamOBP71†	
72		AGAP012714#	Not available	Not available	AgamOBP83	AgamOBP72†	
73		Not available	Not available	Not available	AgamOBP65		No genomic data in VectorBase
74		AGAP008280#	Not available	Not available	AgamOBP69		D7 protein
75		AGAP008281#	Not available	Not available	AgamOBP70		D7 protein
76		AGAP008282#	Not available	Not available	AgamOBP71		D7 protein
77		AGAP008283#	Not available	Not available	AgamOBP72		D7 protein
78		AGAP008284#	Not available	Not available	AgamOBP73		D7 protein
79		AGAP008278#	Not available	Not available	AgamOBP74		D7 protein
80		AGAP008279#	Not available	Not available	AgamOBP75		D7 protein
81		AGAP008279#	Not available	Not available	AgamOBP76		D7 protein
82		AGAP006278#	Not available	Not available	AgamOBP78		D7 protein
83		Not available	Not available	Not available	AgamOBP81		No genomic data in VectorBase

Vectorbase IDs specifically reported in Vieira & Rozas (2011).

† Newly identified OBP genes in Vieira & Rozas (2011) not included in this study.

Supplementary Table 1c. Complete list of OBP genes from *Aedes aegypti* reported in this study. Alongside their identification and chromosomal locations, shown are their phylogenetic clusters and orthologues in *Anopheles gambiae* and *Culex quinquefasciatus*. The 47 OBPs newly identified in this study (AaegOBP67-AaegOBP114) are indicated by an (*). †Two genes AaegOBP42 and AaegOBP63 share 100% sequence identity but are localized on different chromosome segments. We cannot resolve which of these two genes is the true orthologue of AgamOBP48 and CquiOBP106. Two way (1:1) orthologues are indicated by a # sign.

n°	ID	Name	Full length	Super contig	Start	End	cluster	Orthologue in <i>An. gambiae</i>	Orthologue in <i>Culex</i>
1	AAEL009449	AaegOBP1	143	1.397	1059780	1064175	OS-E/OS-F	AgamOBP1	CquiOBP1
2	AAEL006176	AaegOBP2	158	1.193	1418261	1435259	Pbprp1	AgamOBP7	CquiOBP7
3	AAEL013018	AaegOBP3	143	1.776	429862	431937	OS-E/OS-F		
4	AAEL000073	AaegOBP4	146	1.1	4140321	4154466	OBP19a		CquiOBP8
5	AAEL000139	AaegOBP5	269	1.2	651524	652909	mplus6	AgamObp59a	CquiOBP109
6	AAEL000821	AaegOBP6	255	1.17	3417930	3418697	matype4		
7	AAEL000833	AaegOBP7	279	1.17	3935676	3936843	matype4		
8	AAEL001826	AaegOBP8	133	1.43	1541221	1541846	mclassic5	AgamOBP21	CquiOBP23
9	AAEL002596	AaegOBP9	132	1.61	1448858	1449413	mclassic3	AgamOBP23	CquiOBP22
10	AAEL007603	AaegOBP10	140	1.266	1269968	1270519	mclassic6	AgamOBP10	CquiOBP24
11	AAEL002587	AaegOBP11	137	1.61	1518773	1519546	mclassic3	AgamOBP25	CquiOBP19
12	AAEL002617	AaegOBP12	132	1.61	1525746	1526203	mclassic3	AgamOBP28	CquiOBP18
13	AAEL002591	AaegOBP13	132	1.61	1541151	1541767	mclassic4		
14	AAEL002605	AaegOBP14	133	1.61	1560540	1560999	mclassic4		
15	AAEL002598	AaegOBP15	136	1.61	1605563	1618454	mclassic1	AgamOBP64	CquiOBP15
16	AAEL003315	AaegOBP16	269	1.83	2455477	2456452	matype4		
17	AAEL004339	AaegOBP17	138	1.115	975305	1055633	mclassic7		CquiOBP53
18	AAEL004342	AaegOBP18	140	1.115	1056506	1057069	mclassic9	AgamOBP14	
19	AAEL004343	AaegOBP19	145	1.115	1059658	1060207	mclassic9		CquiOBP51
20	AAEL005778	AaegOBP20	166	1.174	827019	827663	Pbprp4	AgamOBP62	CquiOBP14
21	AAEL005770	AaegOBP21	146	1.174	1511030	1512166	Obp99a		
22	AAEL005772	AaegOBP22	138	1.174	1532891	1533675	Obp99a	AgamOBP9	CquiOBP43
23	AAEL006109	AaegOBP23	242	1.189	217057	217969	mplus9		
24	AAEL006108	AaegOBP24	200	1.189	234195	245248	mplus9		
25	AAEL006103	AaegOBP25	200	1.189	2037030	2054778	mplus9	AgamOBP57	CquiOBP104
26	AAEL006106	AaegOBP26	352	1.189	2063847	2065664	mplus9	AgamOBP56	CquiOBP102
27	AAEL000071	AaegOBP27	141	1.1	4056982	4057610	OBP19a	AgamOBP6	CquiOBP13
28	AAEL006393	AaegOBP28	322	1.203	1485149	1497912	matype3		
29	AAEL006387	AaegOBP29	286	1.203	1497040	1521142	matype3	AgamOBP31	
30		AaegOBP30		Previously reported, this gene is not available in VectorBase					

n°	ID	Name	Full length	Super contig	Start	End	cluster	Orthologue in <i>An. gambiae</i>	Orthologue in <i>Culex</i>
31	AAEL006396	AaegOBP31	331	1.203	1526927	1527922	matype3		
32	AAEL006398	AaegOBP32	336	1.203	1537009	1538019	matype3		CquiOBP86
33	AAEL006385	AaegOBP33	313	1.203	1538561	1539571	matype3		
34	AAEL014082	AaegOBP34	149	1.1002	187323	188351	LUSH		
35	AAEL002606	AaegOBP35	131	1.61	1497852	1498861	mclassic3	AgamOBP26	CquiOBP20
36	AAEL008011	AaegOBP36	152	1.294	802739	803197	OS-E/OS-F	AgamOBP15	CquiOBP5
37	AAEL008009	AaegOBP37	148	1.294	823671	828548	OS-E/OS-F		CquiOBP4
38	AAEL008013	AaegOBP38	140	1.294	1185169	1186628	OS-E/OS-F	AgamOBP3	CquiOBP2
39	AAEL006454	AaegOBP39	146	ig1.206	50736	67086	LUSH	AgamOBP4	CquiOBP6
40	AAEL009597	AaegOBP40	291	1.411	257365	258240	matype1	AgamOBP39	CquiOBP84
41	AAEL009599	AaegOBP41	297	1.411	258377	259352	matype1		CquiOBP85
42	AAEL010666	AaegOBP42†	157	1.495	848252	848790	mplus1	†AgamOBP48	†CquiOBP106
43	AAEL010662	AaegOBP43	193	1.495	849614	850509	mplus1		
44	AAEL010718	AaegOBP44	219	1.500	470519	471178	matype4		CquiOBP78
45	AAEL010714	AaegOBP45	260	1.500	485823	486827	matype4		
46	AAEL010872	AaegOBP46	298	1.514	549927	550986	matype4		
47	AAEL011499	AaegOBP47	191	1.584	253363	254080	mplus1		
48	AAEL011494	AaegOBP48	214	1.584	270443	271204	mplus2	AgamOBP46	CquiOBP108
49	AAEL011484	AaegOBP49	187	1.584	357678	358469	mplus4		
50	AAEL011490	AaegOBP50	199	1.584	358749	359468	mplus4		
51	AAEL011487	AaegOBP51	195	1.584	391587	398988	mplus4	AgamOBP60	
52	AAEL011491	AaegOBP52	184	1.584	422989	423783	mplus4		
53	AAEL011482	AaegOBP53	182	1.584	423953	424743	mplus4		CquiOBP110
54	AAEL011481	AaegOBP54	181	1.584	434945	435608	mplus4		
55	AAEL012377	AaegOBP55	151	1.685	122245	142217	OBP19a	AgamOBP20	CquiOBP11
56	AAEL000051	AaegOBP56	115	1.1	4124658	4140143	OBP19a	AgamOBP19	CquiOBP12
57	AAEL000035	AaegOBP57	151	1.1	3668288	3668784	Pbprp2/5	AgamOBP13	CquiOBP28
58	AAEL014430	AaegObp59a	285	1.1115	141516	149446	NOGROUP		
59	AAEL015313	AaegOBP59	166	1.1784	37928	38729	Pbprp4		
60	AAEL015499	AaegOBP60	150	1.2733	6977	7459	OS-E/OS-F	AgamOBP2	CquiOBP3
61	AAEL015554	AaegOBP61	132	1.3221	5993	6448	mclassic4	AgamOBP63	CquiOBP17
62	AAEL015566	AaegOBP62	193	1.3337	1317	2221	mplus1		
63	AAEL015567	AaegOBP63†	157	1.3337	2936	3570	mplus1	†AgamOBP48	†CquiOBP106
64		AaegOBP64		Previously reported, this gene is not available in VectorBase					
65	AAEL002618	AaegOBP65	98	1.61	1577860	1578251	mclassic1	AgamOBP27	CquiOBP16
66		AaegOBP66		Previously reported but same gene as AaegOBP35					

n°	ID	Name	Full length	Super contig	Start	End	cluster	Orthologue in <i>An. gambiae</i>	Orthologue in <i>Culex</i>
67	AAEL011497	AaegOBP67*	158	1.584	380693	381721	mplus4		
68	AAEL011489	AaegOBP68*	189	1.584	360426	380332	mplus4		
69	AAEL000124	AaegOBP69*	180	1.2	673060	674073	mplus7	AgamOBP54	
70	AAEL006105	AaegOBP70*	227	1.189	2070866	2071667	mplus9		CquiOBP103
71	AAEL006094	AaegOBP71*	227	1.189	205937	206741	mplus9		
72	AAEL004729	AaegOBP72*	178	1.128	259634	260619	mplus7		CquiOBP111
73	AAEL004730	AaegOBP73*	168	1.128	267328	268254	mplus7	AgamOBP49	CquiOBP112
74	AAEL011486	AaegOBP74*	193	1.584	304385	305157	mplus5		
75	AAEL011483	AaegOBP75*	193	1.584	321689	322507	mplus5		CquiOBP107
76	AAEL007604	AaegOBP76*	134	1.266	1288889	1290609	mclassic6		
77	AAEL002626	AaegOBP77*	138	1.61	1455363	1456069	mclassic2	AgamOBP24	CquiOBP21
78	AAEL001836	AaegOBP78*	135	1.43	1535067	1535644	minus C		CquiOBP70
79	AAEL007014	AaegOBP79*	154	1.132	2272455	2273324	NOGROUP		
80	AAEL007003	AaegOBP80*	152	1.231	1719941	1720462	NOGROUP		
81	AAEL011730	AaegOBP81*	151	1.606	650035	650560	mclassic8	AgamOBP22	CquiOBP44
82	AAEL014593	AaegOBP82*	145	1.1181	11402	11955	mplus3		CquiOBP101
83	AAEL011416	AaegOBP83*	304	1.579	450825	451739	Obp59a	AgamOBP29	CquiOBP55
84	AAEL000827	AaegOBP84*	287	1.17	3989260	3990220	matype4		
85	AAEL000831	AaegOBP85*	253	1.17	3976929	3977690	matype4		
86	AAEL004856	AaegOBP86*	255	1.132	2272456	2273324	matype4		
87	AAEL003511	AaegOBP87*	265	1.89	1577417	1578335	matype4		CquiOBP76
88	AAEL010874	AaegOBP88*	299	1.514	517957	519061	matype4		
89	AAEL000377	AaegOBP89*	315	1.6	3059016	3060090	matype2		
90	AAEL013719	AaegOBP90*	202	1.902	36629	37757	matype2		
91	AAEL013720	AaegOBP91*	273	1.902	26444	29202	matype2		CquiOBP87
92	AAEL000318	AaegOBP92*	294	1.6	3110479	3111363	matype2		
93	AAEL000319	AaegOBP93*	287	1.6	3087173	3088138	matype2		
94	AAEL000344	AaegOBP94*	312	1.6	3052765	3053725	matype2		
95	AAEL000350	AaegOBP95*	294	1.6	3114188	3116137	matype2		
96	AAEL000796	AaegOBP96*	305	1.17	3920770	3922193	matype4		
97	AAEL000835	AaegOBP97*	260	1.17	3900849	3901700	matype4		
98	AAEL001174	AaegOBP98*	586	1.24	2847433	2850832	matype2		CquiOBP95
99	AAEL001179	AaegOBP99*	332	1.24	2705024	2711778	matype2		
100	AAEL003513	AaegOBP100*	278	1.89	1529116	1529952	matype4		CquiOBP77
101	AAEL003525	AaegOBP101*	370	1.89	1489315	1498788	matype4		
102	AAEL003538	AaegOBP102*	291	1.89	1554285	1555425	matype4	AgamOBP32	

n°	ID	Name	Full length	Super contig	Start	End	cluster	Orthologue in <i>An. gambiae</i>	Orthologue in <i>Culex</i>
103	AAEL010875	AaegOBP103*	299	1.514	538264	539163	matype4		
104	AAEL004516	AaegOBP104*	295	1.122	2397633	2398520	matype2		
105	AAEL001189	AaegOBP105*	309	1.24	2833159	2834187	matype2		
106	AAEL001153	AaegOBP106*#	302	1.24	2825960	2826974	matype2	AgamOBP34/37#	CquiOBP97#
107	AAEL014876	AaegOBP107*	299	1.1319	90885	91826	matype2		CquiOBP94
108	AAEL014874	AaegOBP108*	278	1.1319	77202	78116	matype2		
109	AAEL009433	AaegOBP109*	278	1.396	502699	503646	matype2		
110	AAEL014431	AaegOBP110*	268	1.1115	149798	150604	NOGROUP		
111	AAEL003311	AaegOBP111*	347	1.83	658245	668798	matype4		
112	AAEL000837	AaegOBP112*	306	1.17	3960390	3961459	matype4		
113	AAEL008640	AaegOBP113*	274	1.338	484248	485189	NOGROUP		
114	AAEL001487	AaegOBP114*	305	1.34	1695923	1698069	NOGROUP		

*Newly discovered genes in this study.

Only a two way (1:1) orthology has been established for these genes and not a three way (1:1:1) orthology.

†Two genes *AaegOBP42* and *AaegOBP63* share 100% sequence identity but are localized on different chromosome segments. We cannot resolve for these two genes which is the true orthologue of *AgamOBP48* and *CquiOBP106*.

Supplementary Table 1d. Complete list of OBP genes from *Culex quinquefasciatus* Reported in this study. Alongside their identification and chromosomal locations, shown are their phylogenetic clusters and orthologues in *Anopheles gambiae* and *Aedes aegypti*. †The last two genes were identified by Vieira & Rozas (2011) and were added to this Table at the last moment after renaming them CquiOBP113 and CquiOBP114. Two way (1:1) orthologues are indicated by a # sign.

n°	ID	Name	Full length	Super contig	Start	End	Cluster	Orthologue in <i>An. gambiae</i>	Orthologue in <i>Ae. aegyptii</i>
1	CPIJ007604	CquiOBP1	149	3.150	170719	174721	OS-E/OS-F	AgamOBP1	AaegOBP1
2	CPIJ007617	CquiOBP2	146	3.150	672931	673546	OS-E/OS-F	AgamOBP3	AaegOBP38
3	CPIJ007611	CquiOBP3	147	3.150	540281	542064	OS-E/OS-F	AgamOBP2	AaegOBP60
4	CPIJ001730	CquiOBP4	150	3.25	734060	734572	OS-E/OS-F		AaegOBP37
5	CPIJ007608	CquiOBP5	143	3.150	516885	517412	OS-E/OS-F	AgamOBP15	AaegOBP36
6	CPIJ008793	CquiOBP6	89	3.206	489697	490937	LUSH	AgamOBP4	AaegOBP39
7	CPIJ001365	CquiOBP7	136	3.18	1720262	1721216	Pbprp1	AgamOBP7	AaegOBP2
8	CPIJ009568	CquiOBP8	144	3.240	122626	123234	OBP19a		AaegOBP4
9	CPIJ016948	CquiOBP9	139	3.865	41129	46297	OBP19a		
10	CPIJ013976	CquiOBP10	132	3.550	256165	256681	OBP19a		
11	CPIJ006551	CquiOBP11	144	3.121	270272	277928	OBP19a	AgamOBP20	AaegOBP55
12	CPIJ016949	CquiOBP12	121	3.865	46518	47165	OBP19a	AgamOBP19	AaegOBP56
13	CPIJ016952	CquiOBP13	143	3.865	54944	61815	OBP19a	AgamOBP6	AaegOBP27
14	CPIJ009586	CquiOBP14	170	3.240	569948	574407	Pbprp4	AgamOBP62	AaegOBP20
15	CPIJ012714	CquiOBP15	141	3.424	103588	109982	mclassic1	AgamOBP64	AaegOBP15
16	CPIJ012715	CquiOBP16	134	3.424	112183	112979	mclassic1	AgamOBP27	AaegOBP65
17	CPIJ012716	CquiOBP17	132	3.424	113896	114578	mclassic4	AgamOBP63	AaegOBP61
18	CPIJ012717	CquiOBP18	132	3.424	122946	123411	mclassic3	AgamOBP28	AaegOBP12
19	CPIJ012718	CquiOBP19	139	3.424	131078	131864	mclassic3	AgamOBP25	AaegOBP11
20	CPIJ012719	CquiOBP20	131	3.424	135879	136509	mclassic3	AgamOBP26	AaegOBP35
21	CPIJ012720	CquiOBP21	139	3.424	171439	171968	mclassic2	AgamOBP24	AaegOBP77
22	CPIJ012721	CquiOBP22	131	3.424	172603	173060	mclassic3	AgamOBP23	AaegOBP9
23	CPIJ001876	CquiOBP23	136	3.26	255589	259525	mclassic5	AgamOBP21	AaegOBP8
24	CPIJ014525	CquiOBP24	137	3.561	24869	25524	mclassic6	AgamOBP10	AaegOBP10
25	CPIJ010723	CquiOBP25	121	3.286	224289	224718	Pbprp2/5		
26	CPIJ010724	CquiOBP26	119	3.286	228005	228420	Pbprp2/5		
27	CPIJ010728	CquiOBP27	126	3.286	489935	490384	Pbprp2/5		
28	CPIJ016965	CquiOBP28	150	3.865	148161	148975	Pbprp2/5	AgamOBP13	AaegOBP57
29	CPIJ016966	CquiOBP29	130	3.865	149508	150489	Pbprp2/5		
30	CPIJ016967	CquiOBP30	143	3.865	154625	155111	Pbprp2/5		
31	CPIJ008285	CquiOBP31	124	3.167	404302	404732	Pbprp2/5		

Supplementary Table 2.1d

n°	ID	Name	Full length	Super contig	Start	End	Cluster	Orthologue in <i>An. gambiae</i>	Orthologue in <i>Ae. aegyptii</i>
32	CPIJ016479	CquiOBP32	126	3.770	2731	3167	Pbprp2/5		
33	CPIJ019607	CquiOBP33	124	3.1894	15149	15587	Pbprp2/5		
34	CPIJ019608	CquiOBP34	116	3.1894	29115	29465	Pbprp2/5		
35	CPIJ019609	CquiOBP35	126	3.1894	31188	31622	Pbprp2/5		
36	CPIJ019610	CquiOBP36	146	3.1894	41408	41883	Pbprp2/5		
37	CPIJ007931	CquiOBP37	135	3.181	460064	466993	Pbprp2/5		
38	CPIJ007932	CquiOBP38	137	3.181	467058	467528	Pbprp2/5		
39	CPIJ007933	CquiOBP39	126	3.181	481658	482092	Pbprp2/5		
40	CPIJ007934	CquiOBP40	107	3.181	487383	487920	Pbprp2/5		
41	CPIJ007935	CquiOBP41	98	3.181	488157	488453	Pbprp2/5		
42	CPIJ007936	CquiOBP42	111	3.181	492753	493384	Pbprp2/5		
43	CPIJ017326	CquiOBP43	138	3.984	153967	154634	Obp99a	AgamOBP9	AaegOBP22
44	CPIJ009937	CquiOBP44	147	3.265	418539	421106	mclassic8	AgamOBP22	AaegOBP81
45		CquiOBP45		Reported in previous paper, not available in VectorBase					
46	CPIJ010782	CquiOBP46	150	3.315	176953	177463	mclassic9		
47		CquiOBP47		Reported in previous paper, not available in VectorBase					
48		CquiOBP48		Reported in previous paper, not available in VectorBase					
49		CquiOBP49		Reported in previous paper, not available in VectorBase					
50		CquiOBP50		Reported in previous paper, not available in VectorBase					
51	CPIJ010787	CquiOBP51	144	3.315	189941	190471	mclassic9		AaegOBP19
52	CPIJ010788	CquiOBP52	143	3.315	190549	191091	mclassic9		
53	CPIJ010789	CquiOBP53	145	3.315	191345	193026	mclassic7		AaegOBP17
54	CPIJ007937	CquiOBP54*	170	3.181	496322	497393	Pbprp2/5		
55	CPIJ010367	CquiOBP55*	235	3.273	146471	147364	Obp59a	AgamOBP29	AaegOBP83
56	CPIJ010729	CquiOBP56*	214	3.286	493370	501823	Pbprp2/5		
57	CPIJ016951	CquiOBP57*	126	3.865	50440	50879	OBP19a		
58	CPIJ007609	CquiObp59a*	141	3.150	526150	526551	OS-E/OS-F		
59	CPIJ001871	CquiOBP59*	113	3.26	242975	243783	minus C		
60	CPIJ012786	CquiOBP60*	138	3.443	357791	358353	minus C		
61	CPIJ015943	CquiOBP61*	138	3.727	98658	99154	minus C		
62	CPIJ015944	CquiOBP62*	120	3.727	99851	104141	minus C		
63	CPIJ016343	CquiOBP63*	181	3.758	25841	28252	mclassic9	AgamOBP14	
64	CPIJ004145	CquiOBP64*	206	3.64	238526	238997	minus C		
65	CPIJ001875	CquiOBP65*	136	3.26	253862	254332	minus C		
66	CPIJ001865	CquiOBP66*	136	3.26	220863	221383	minus C		
67	CPIJ017432	CquiOBP67*	130	3.930	38452	43014	minus C		

Supplementary Table 2.1d

n°	ID	Name	Full length	Super contig	Start	End	Cluster	Orthologue in <i>An. gambiae</i>	Orthologue in <i>Ae. aegyptii</i>
68	CPIJ001873	CquiOBP68*	119	3.26	248382	249594	minus C		
69	CPIJ001874	CquiOBP69*	137	3.26	249635	250213	minus C		
70	CPIJ001872	CquiOBP70*	122	3.26	247521	248064	minus C		AaegOBP78
71	CPIJ001870	CquiOBP71*	136	3.26	234093	234614	minus C		
72	CPIJ001867	CquiOBP72*	134	3.26	226275	226723	minus C		
73	CPIJ001869	CquiOBP73*	98	3.26	232747	233357	minus C		
74	CPIJ001868	CquiOBP74*	132	3.26	226902	227400	minus C		
75	CPIJ008157	CquiOBP75*	128	3.183	230895	231773	matype4		
76	CPIJ008158	CquiOBP76#*	292	3.183	231998	233839	matype4	AgamOBP43#	AaegOBP87#
77	CPIJ008159	CquiOBP77*	521	3.183	234041	234919	matype4		AaegOBP100
78	CPIJ008155	CquiOBP78*	292	3.183	226602	227315	matype4		AaegOBP44
79	CPIJ008156	CquiOBP79*	237	3.183	227485	230705	matype4		
80	CPIJ008160	CquiOBP80*	250	3.183	235156	236019	matype4		
81	CPIJ008161	CquiOBP81*	287	3.183	237704	238659	matype4		
82	CPIJ008154	CquiOBP82*	309	3.183	225309	225992	matype4		
83	CPIJ000653	CquiOBP83*	227	3.7	1883021	1884070	matype4		
84	CPIJ015732	CquiOBP84*	349	3.670	26079	26960	matype1	AgamOBP39	AaegOBP40
85	CPIJ015733	CquiOBP85*	293	3.670	27043	27906	matype1		AaegOBP41
86	CPIJ009038	CquiOBP86*	287	3.216	624714	626149	matype3		AaegOBP32
87	CPIJ003865	CquiOBP87*	313	3.54	1105231	1106267	matype2		AaegOBP91
88	CPIJ003863	CquiOBP88*	307	3.54	1100660	1101583	matype2		
89	CPIJ003866	CquiOBP89*	307	3.54	1107587	1108531	matype2		
90	CPIJ003867	CquiOBP90*	314	3.54	1112050	1113000	matype2		
91	CPIJ001690	CquiOBP91*	316	3.19	1064263	1065189	matype2		
92	CPIJ017342	CquiOBP92*	308	3.908	23349	24732	matype3		
93	CPIJ017170	CquiOBP93*	353	3.874	68127	69047	matype2		
94	CPIJ017166	CquiOBP94*	306	3.874	53265	54562	matype2		AaegOBP107
95	CPIJ017167	CquiOBP95*	303	3.874	54671	57028	matype2		AaegOBP98
96	CPIJ017164	CquiOBP96*	506	3.874	42543	45345	matype2		
97	CPIJ017165	CquiOBP97*	333	3.874	45828	48052	matype2		AaegOBP106
98	CPIJ017163	CquiOBP98*	309	3.874	40347	41997	matype2		
99	CPIJ017169	CquiOBP99*	400	3.874	66534	67385	matype2		
100	CPIJ017168	CquiOBP100*	274	3.874	59822	60630	matype2		
101	CPIJ008979	CquiOBP101*	251	3.212	2741	3415	mplus3		AaegOBP82
102	CPIJ004634	CquiOBP102*	242	3.70	262125	262974	mplus9	AgamOBP56	AaegOBP26
103	CPIJ004635	CquiOBP103*	228	3.70	268367	269111	mplus9		AaegOBP70

n°	ID	Name	Full length	Super contig	Start	End	Cluster	Orthologue in <i>An. gambiae</i>	Orthologue in <i>Ae. aegyptii</i>
104	CPIJ004630	CquiOBP104*	204	3.70	243833	244540	mplus9	AgamOBP57	AaegOBP25
105	CPIJ002106	CquiOBP105*	127	3.21	701497	701945	mplus1		
106	CPIJ002105	CquiOBP106*	194	3.21	699415	700558	mplus1	AgamOBP48	AaegOBP42
107	CPIJ002109	CquiOBP107*	195	3.21	708381	709996	mplus5		AaegOBP75
108	CPIJ002108	CquiOBP108*	205	3.21	705497	706243	mplus2	AgamOBP46	AaegOBP48
109	CPIJ006608	CquiOBP109*	218	3.130	754816	760061	mplus6	AgamObp59a	AaegOBP5
110	CPIJ002111	CquiOBP110#*	191	3.21	713441	714192	mplus4	AgamOBP60#	AaegOBP53#
111	CPIJ008867	CquiOBP111#*	172	3.219	551750	552504	mplus7	AgamOBP54#	AaegOBP72#
112	CPIJ008868	CquiOBP112*	175	3.219	554916	555746	mplus7	AgamOBP49	AaegOBP73
113	CPIJ017524	CquiOBP113†	178	3.978	126190	127209	nd		
114	CPIJ007337	CquiOBP114†	194	3.157	294891	306059	nd	AGAP006368	

*Newly discovered genes in this study.

Only a two way (1:1) orthology has been established for these genes and not a three way (1:1:1) orthology.

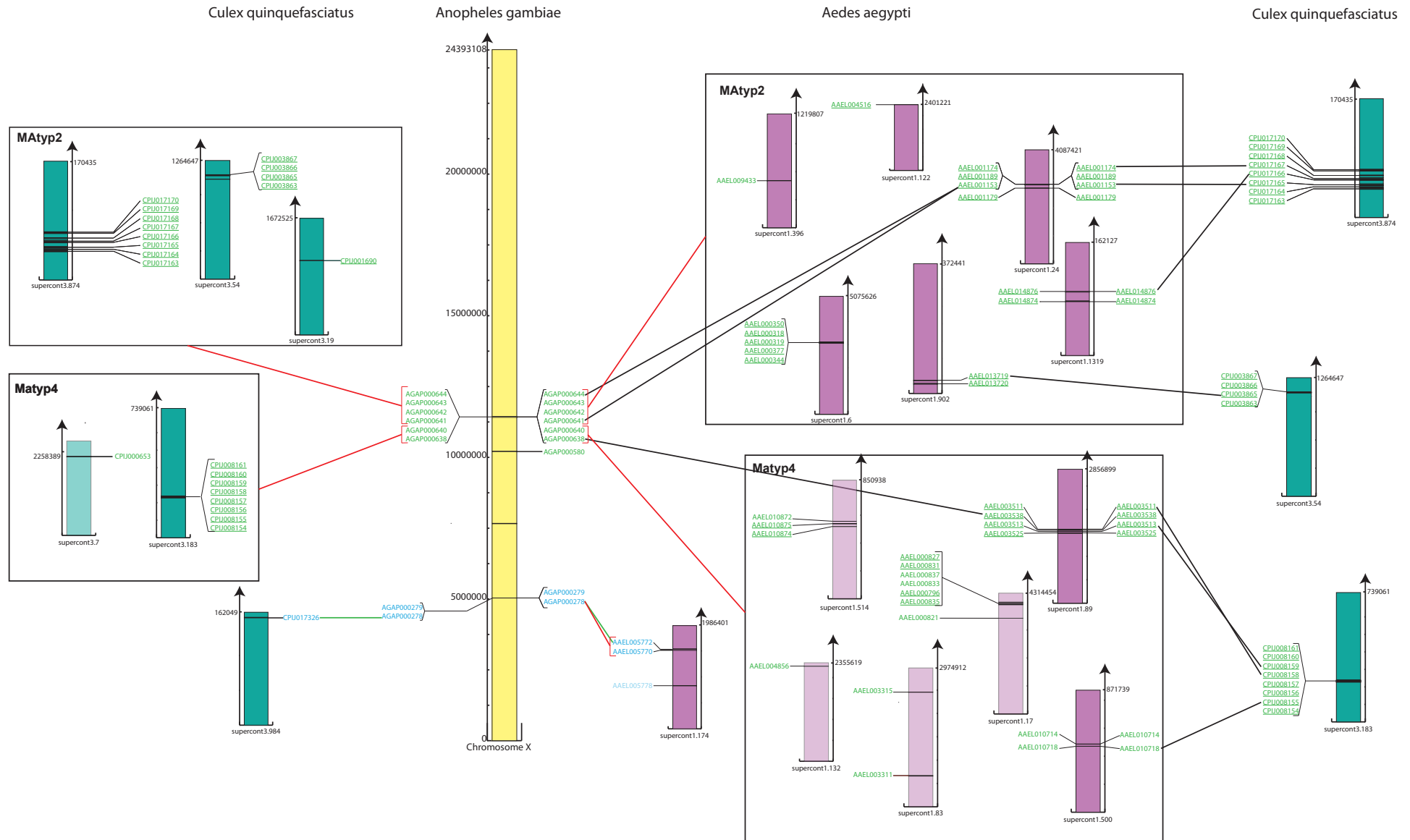
nd : not determined

† New genes recently reported by Vieira & Rozas (2011)

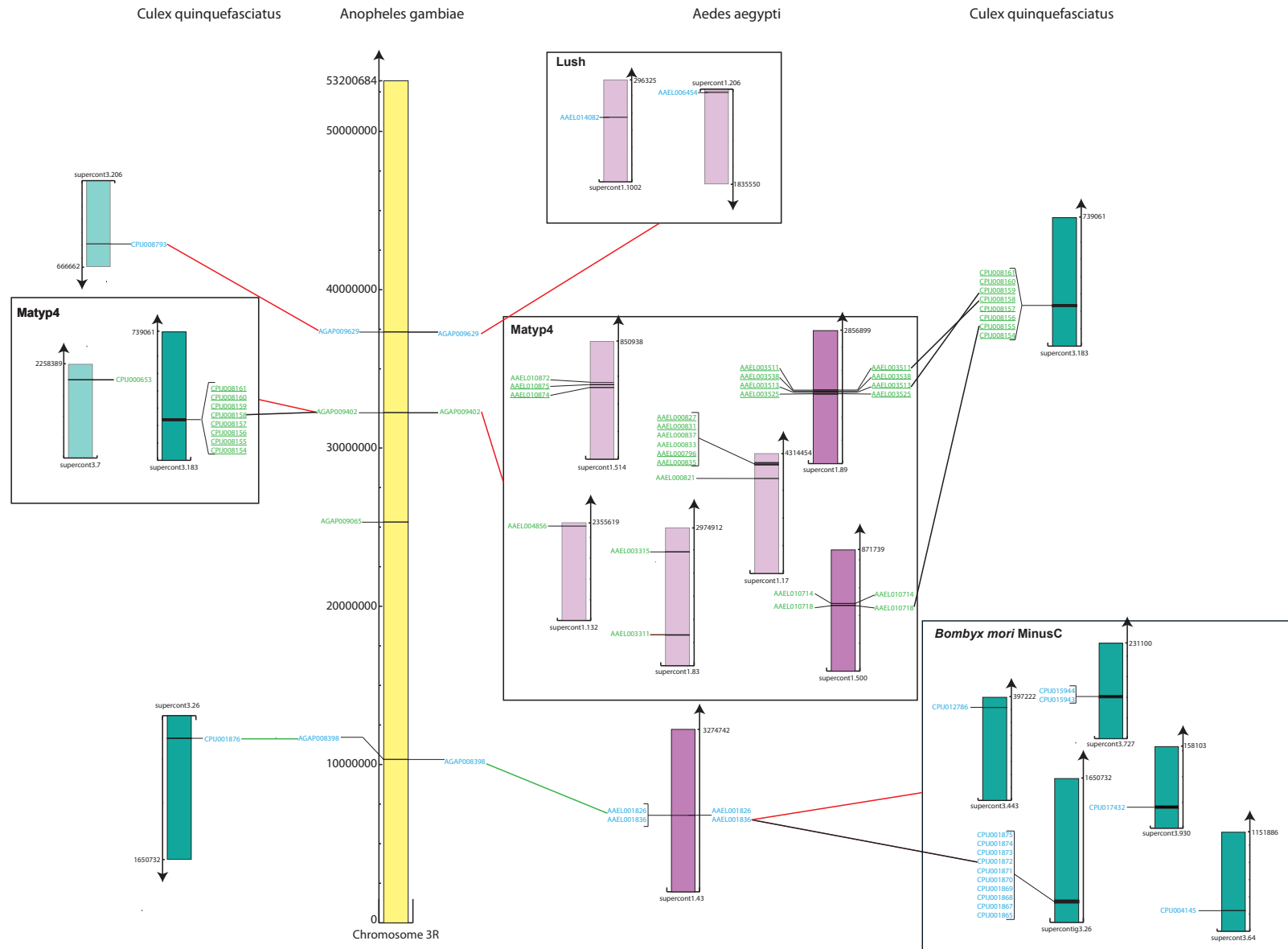
Supplementary Table 2. Syntheny between chromosomes between the four dipterian species *Drosophila melanogaster*; *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* (Arensburger *et al*, 2010).

Drosophila	Anopheles	Aedes	Culex
X	X	1p	1p
		1q	1q
2L	3R	2q	2p
2R	3L	3q	2q
3L	2L	2p	3p & 2q
3R	2R	3p	3q & 1q

Reference : Peter Arensburger, *et al*. Sequencing of *Culex quinquefasciatus* Establishes a Platform for Mosquito Comparative Genomics. *Science* **330**:86-88, (2010).

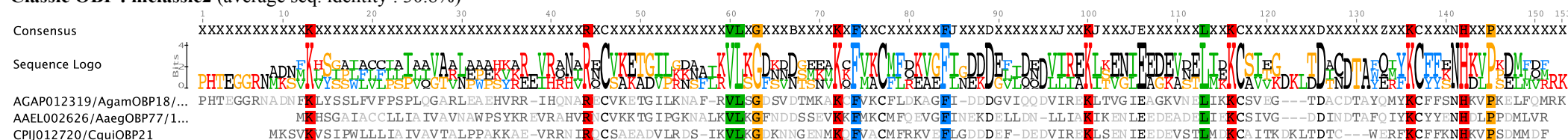
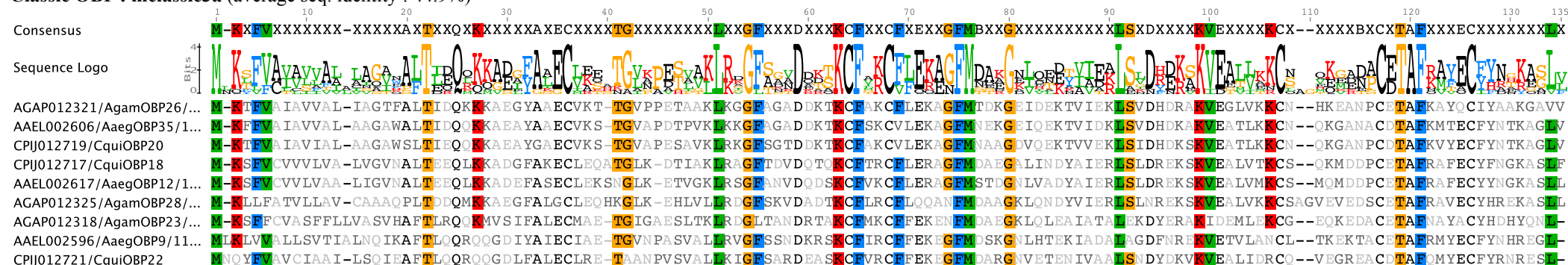


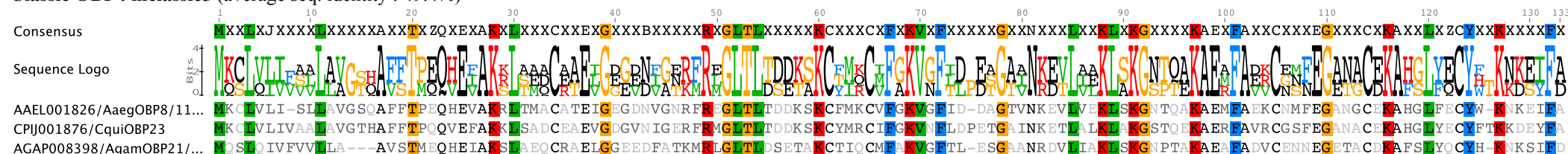
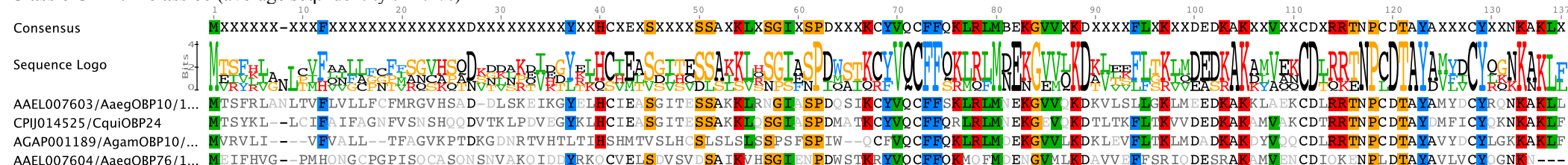
Supplementary figure 1a. Analysis of OBP genes distribution on X chromosome (yellow bar) of *Anopheles gambiae* and their corresponding orthologues & paralogues on *Aedes aegypti* (purple bars) and *Culex quinquefasciatus* (green) supercontigs. See additional legend for more details.

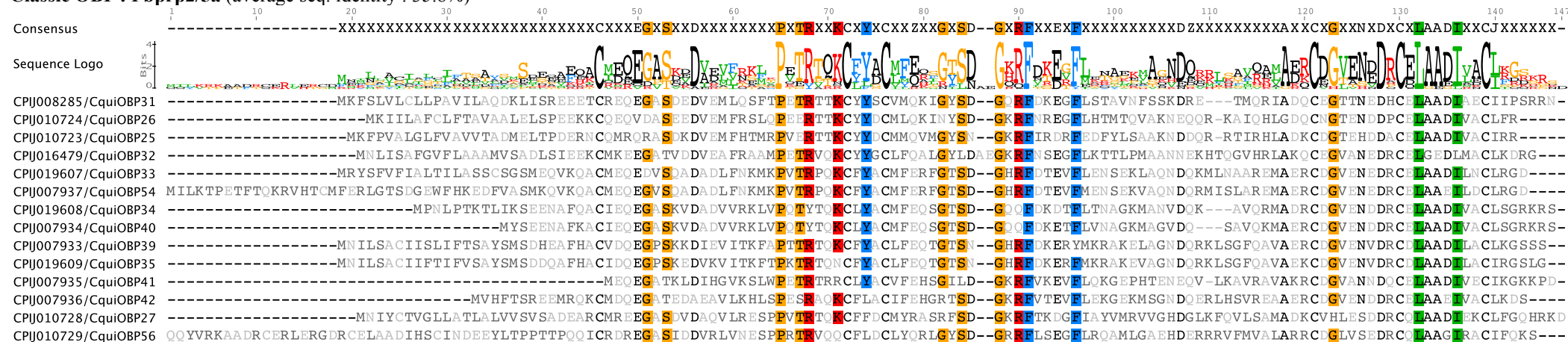
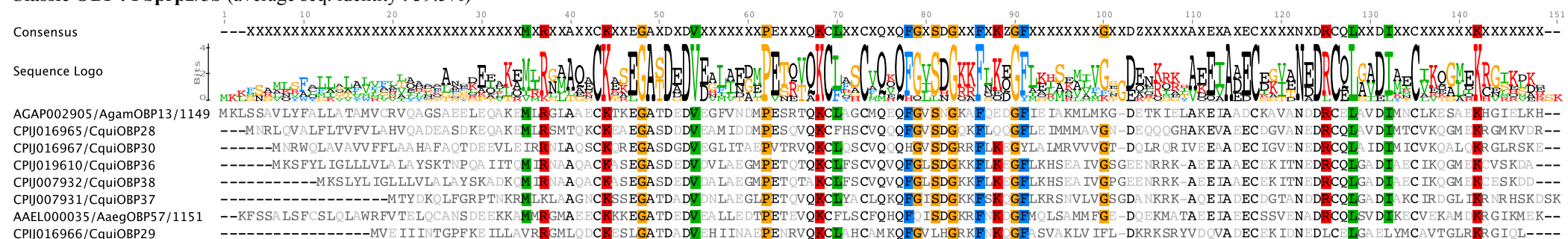
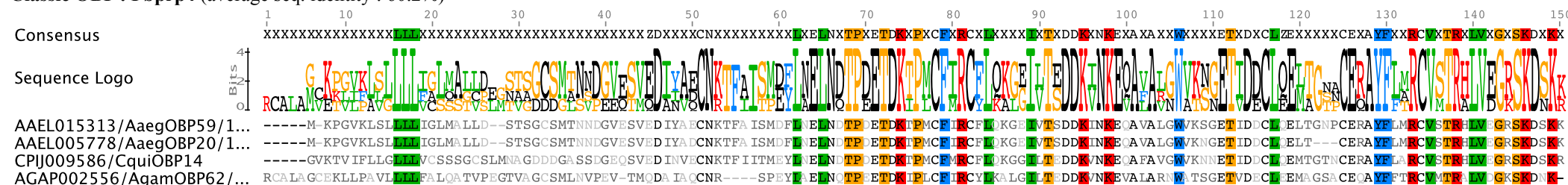


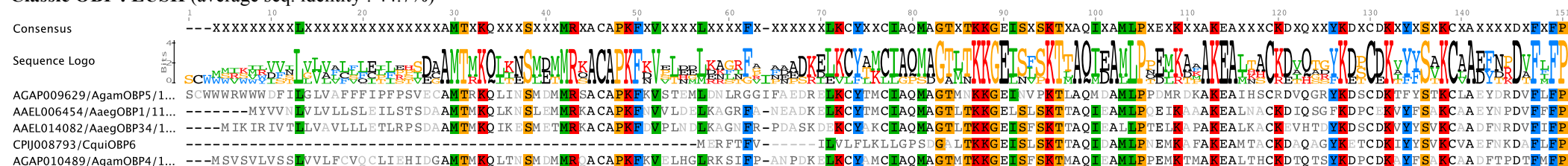
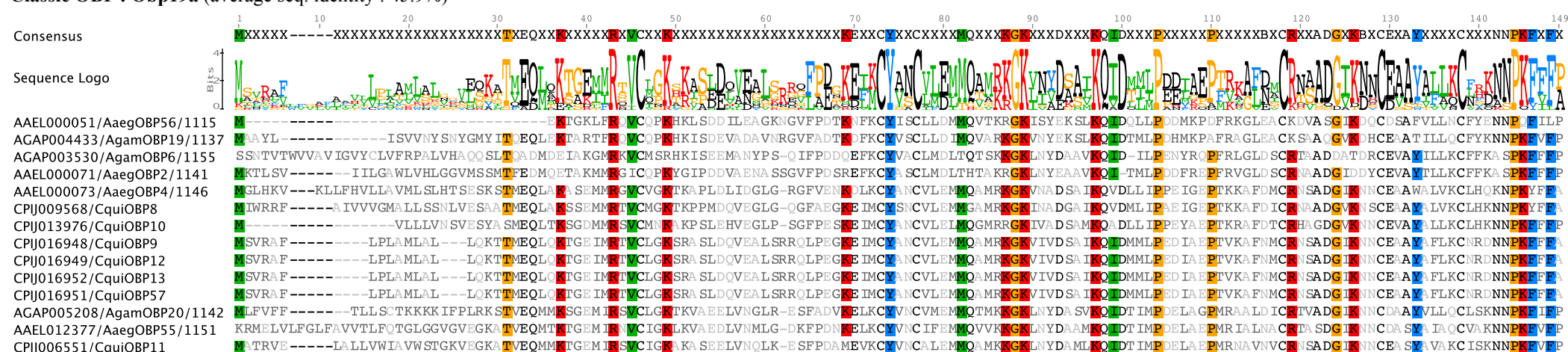
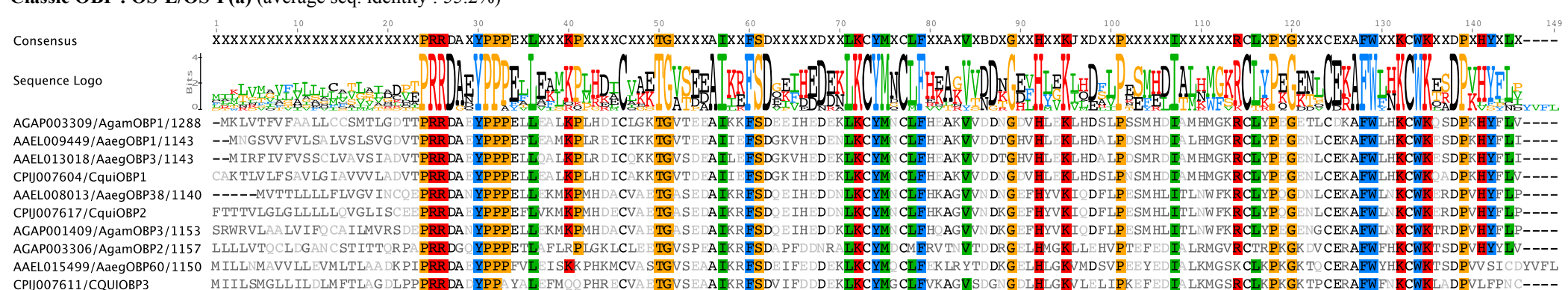
Supplementary figure 1e. Analysis of OBP genes distribution on 3R chromosome (yellow bar) of *Anopheles gambiae* and their corresponding orthologues & paralogues on *Aedes aegypti* (purple bars) and *Culex quinquefasciatus* (green bars) supercontigs. See additional legend for more details.

Additional legend to supplementary Figures 1a-e. Analysis of OBP genes distributions in *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus* genomes. Genes from the subfamilies of the OBP group are colored differently: *Classic* OBPs genes are printed in blue, *Atypical* OBPs in green and *PlusC* in red. The *An. gambiae* chromosomes are in yellow and are centrally located in the diagrams. The *Ae. aegypti* and *C. quinquefasciatus* super contigs are featured in purple and green respectively. Orthology between OBP genes was mainly established using the reverse blast hit (*rbh*) methodology (see materials and method for details). Paralogous relationships were confirmed through examination of the corresponding entries in the *inParanoid* database. Two-way orthologous relationships *i.e* only between genes in two genomes are connected with black lines while three-way orthologous relationships are featured using green lines. Red lines indicate inparalogous links between the connected sets of genes. The contigs from *C. quinquefasciatus* or *A. aegypti* are grouped in a square when all the enclosed OBP genes are from the same phylogenetic subcluster and are inparalogues, except for genes or contigs that are colored in semi-transparency : in these cases, the genes or contigs are displayed to recall their existence in the given cluster but do not share inparalogy relationship with the other enclosed OBP genes ; they might be orthologous to other genes in other chromosomal location. Underlined genes are newly identified genes in this work.

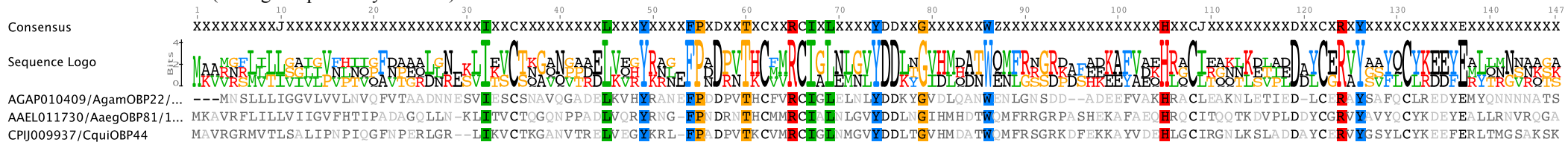
Classic OBP : mclassic1 (average seq. identity : 39.8%)**Classic OBP : mclassic2** (average seq. identity : 30.8%)**Classic OBP : mclassic3a** (average seq. identity : 44.9%)**Classic OBP : mclassic3b** (average seq. identity : 56.6%)**Supplementary Figure 2a.** Protein sequence alignments of OBPs that belongs to the different clusters of the *Classic* subfamily. See additional legend for details.

Classic OBP : mclassic4 (average seq. identity : 77.3%)**Classic OBP : mclassic5** (average seq. identity : 49.4%)**Classic OBP : mclassic6** (average seq. identity : 41.4%)**Classic OBP : Obp59a** (average seq. identity : 56.0%)**Classic OBP : Pbprp1** (average seq. identity : 53.6%)Supplementary Figure 2a. Protein sequence alignments of OBPs that belongs to the different clusters of the *Classic* subfamily. See additional legend for details.

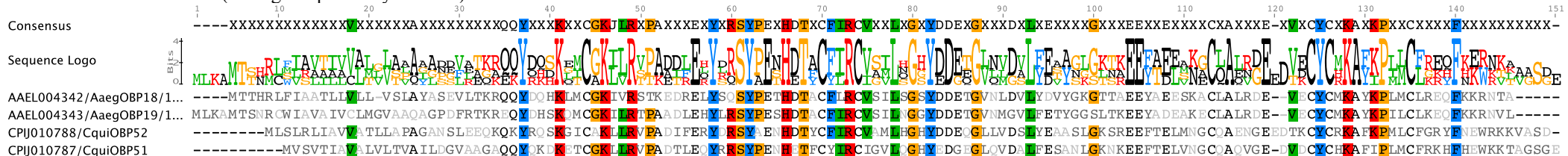
Classic OBP : Pbprp2/5a (average seq. identity : 35.8%)**Classic OBP : Pbprp2/5b** (average seq. identity : 39.5%)**Classic OBP : Pbprp4** (average seq. identity : 60.2%)Supplementary Figure 2a. Protein sequence alignments of OBPs that belongs to the different clusters of the *Classic* subfamily. See additional legend for details.

Classic OBP : LUSH (average seq. identity : 44.7%)**Classic OBP : Obp19a (average seq. identity : 45.9%)****Classic OBP : OS-E/OS-F(a) (average seq. identity : 55.2%)****Supplementary Figure 2a.** Protein sequence alignments of OBPs that belongs to the different clusters of the *Classic* subfamily. See additional legend for details.

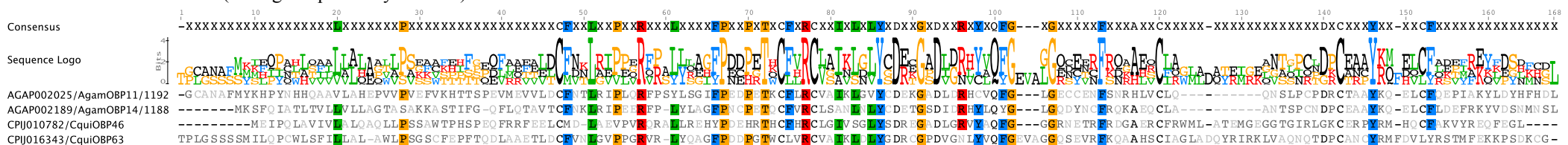
Classic OBP : mclassic8 (average seq. identity : 30.5%)



Classic OBP : mclassic9a (average seq. identity : 40.6%)

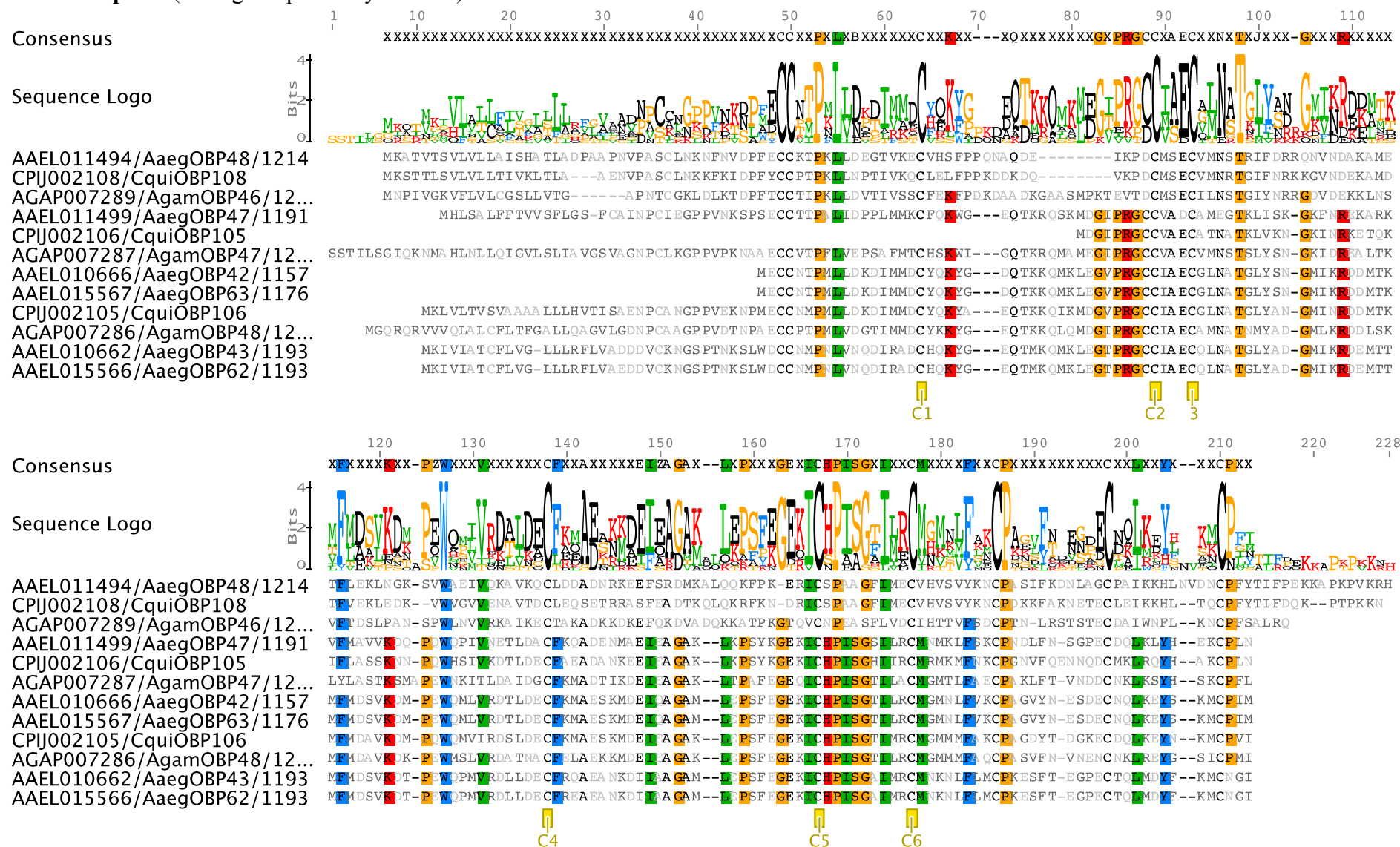


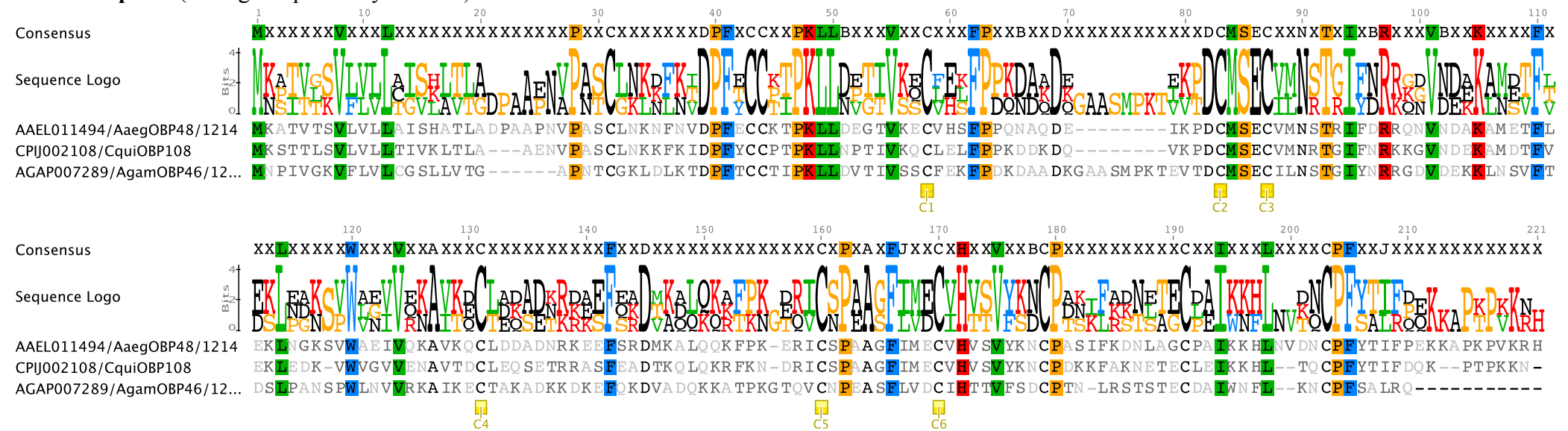
Classic OBP : mclassic9b (average seq. identity : 25.1%)

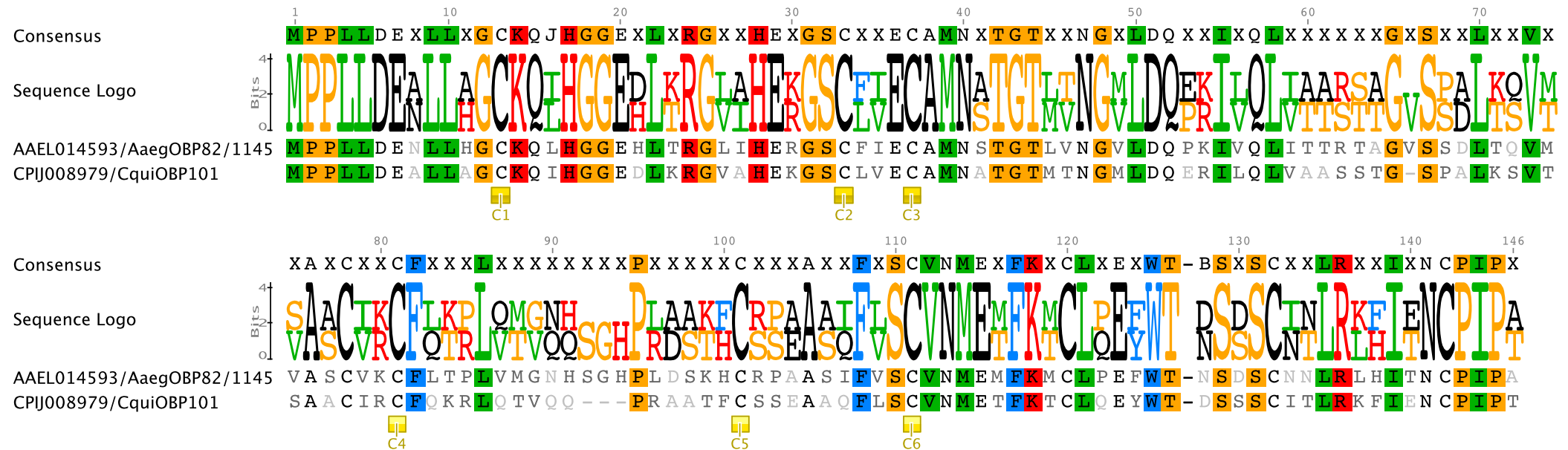


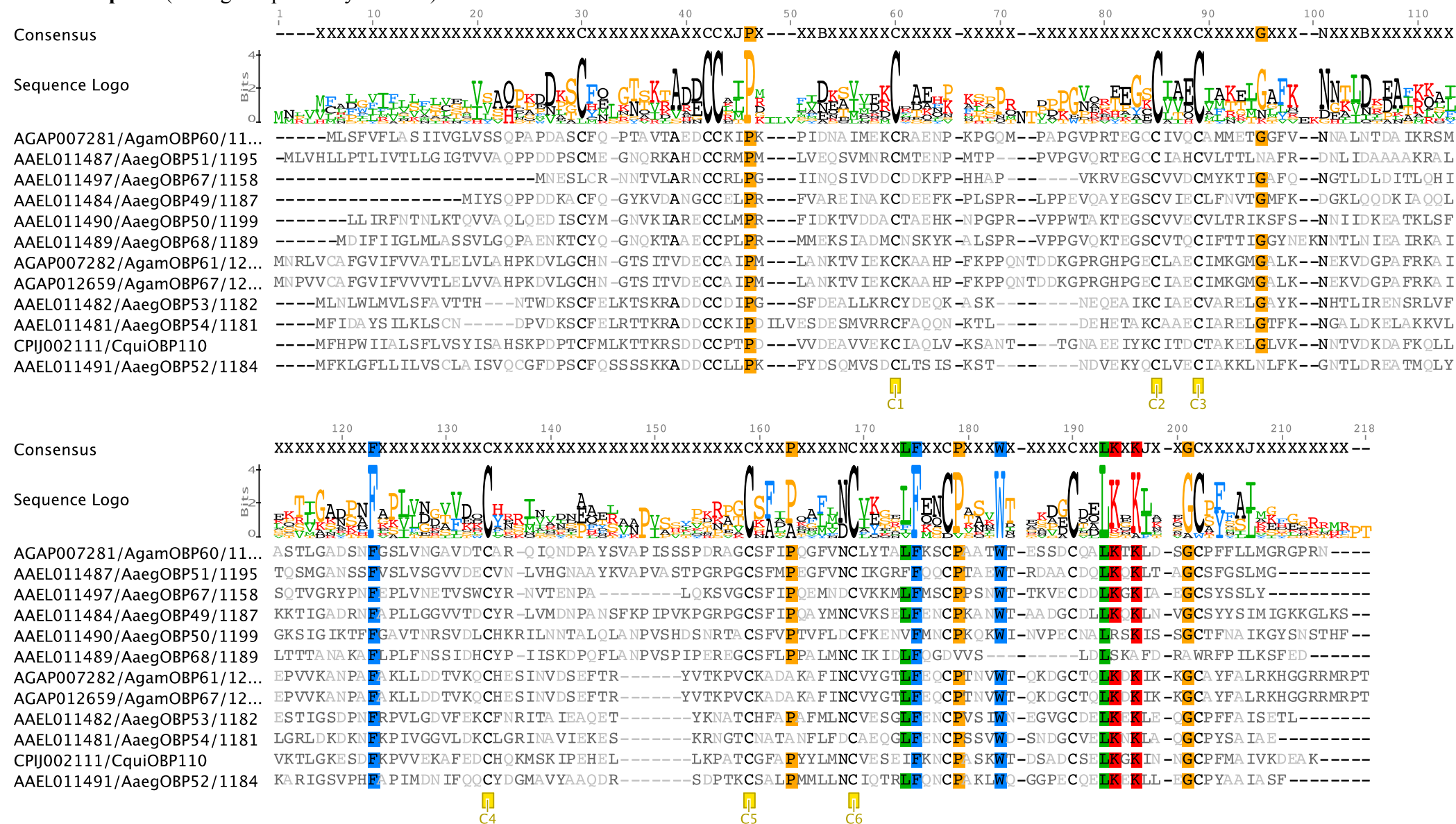
Additional legend to Figure 2a.

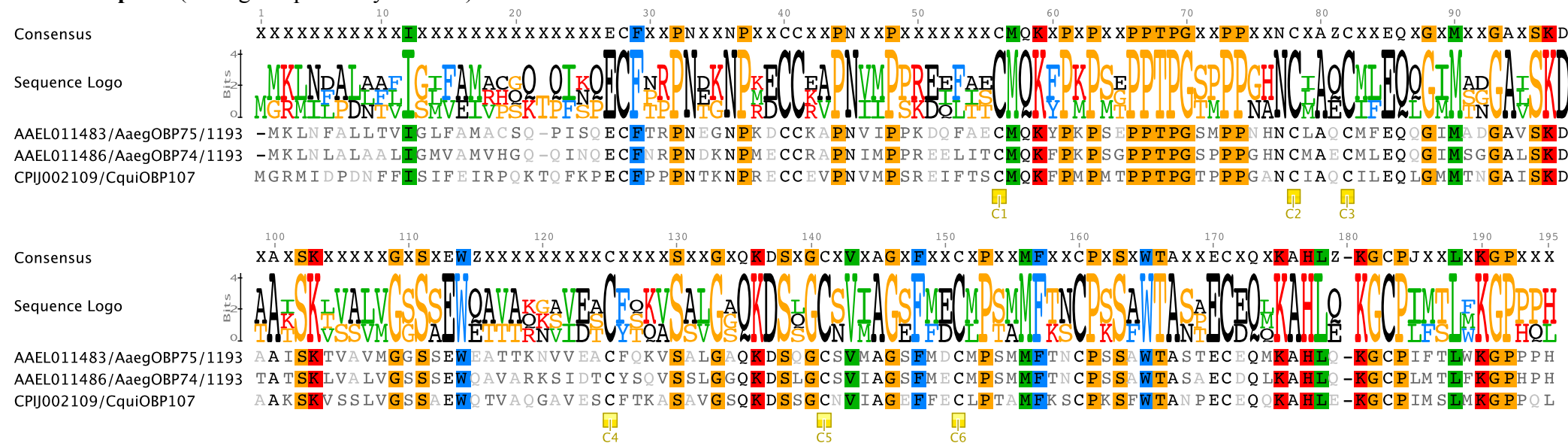
Shown are the alignments for the *Classic* OBPs from the three mosquito genomes *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*. These were obtained after aligning the sequences with a structure-based profile using ClustalX software. The global *Classic* OBP alignment was splitted into different parts corresponding to the phylogenetic clusters that is established in this study (see Figures 4 and 5). The residues that are highly conserved (with 75% or more degree of conservation) are highlighted in the sequences and in the sequence logos above the alignment. The consensus sequence is also featured on top of each alignment. The average pairwise sequence identities within each cluster are indicated. These diagrams were generated by the Geneious software.

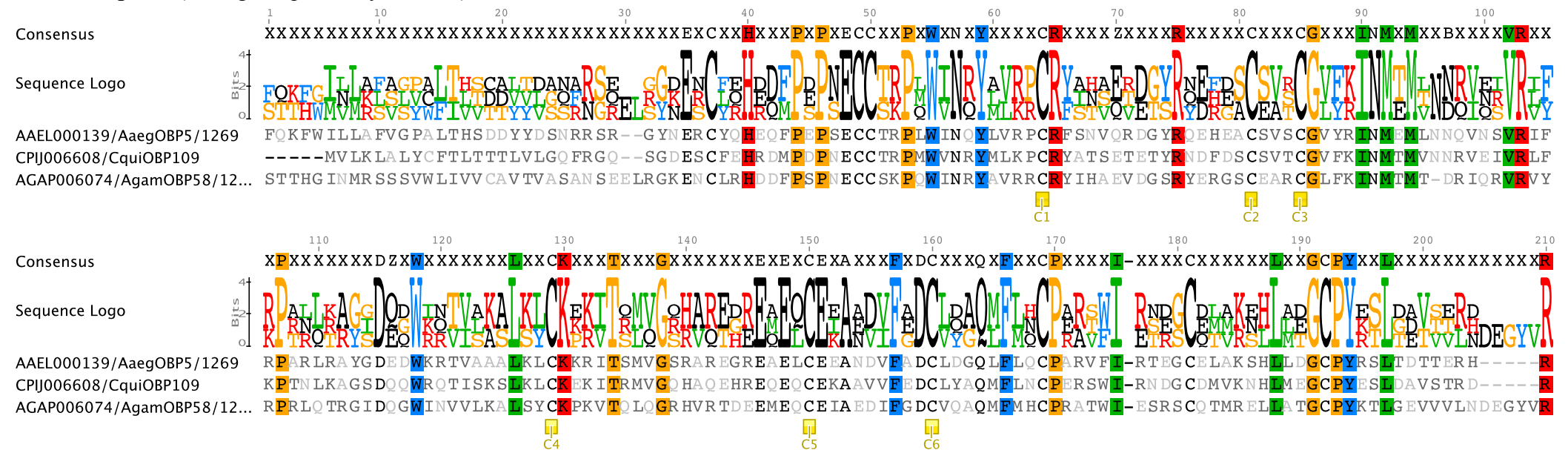
Plus C : *mplus1* (average seq. identity : 41.5%)Supplementary Figure 2b. Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

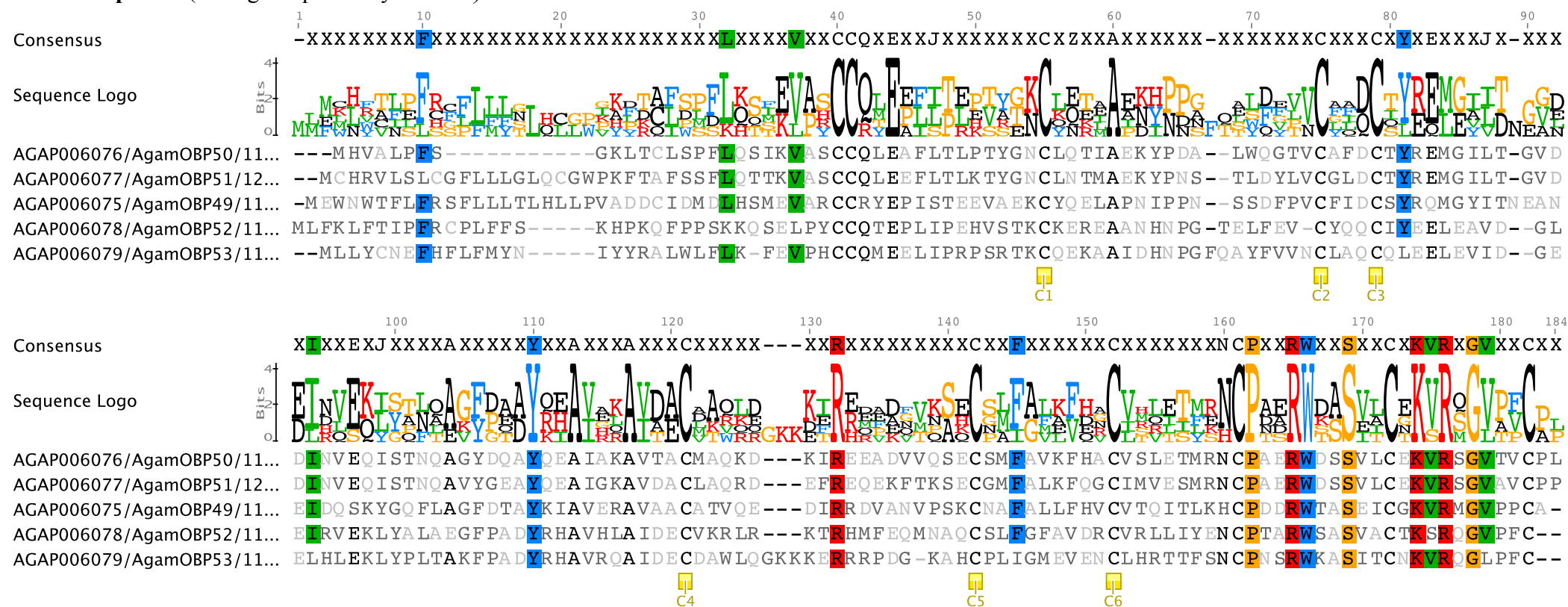
Plus C : mplus2 (average seq. identity : 40.4%)**Supplementary Figure 2b.** Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

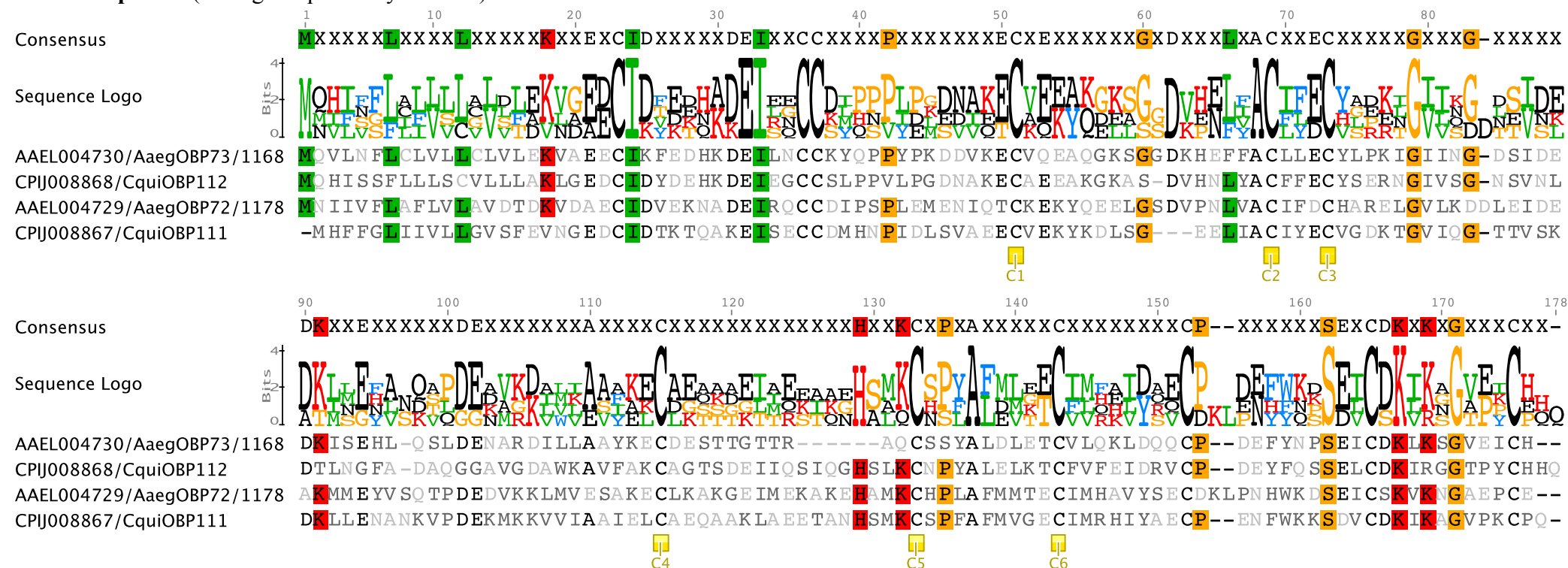
Plus C : mplus3 (average seq. identity : 53.8%)**Supplementary Figure 2b.** Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

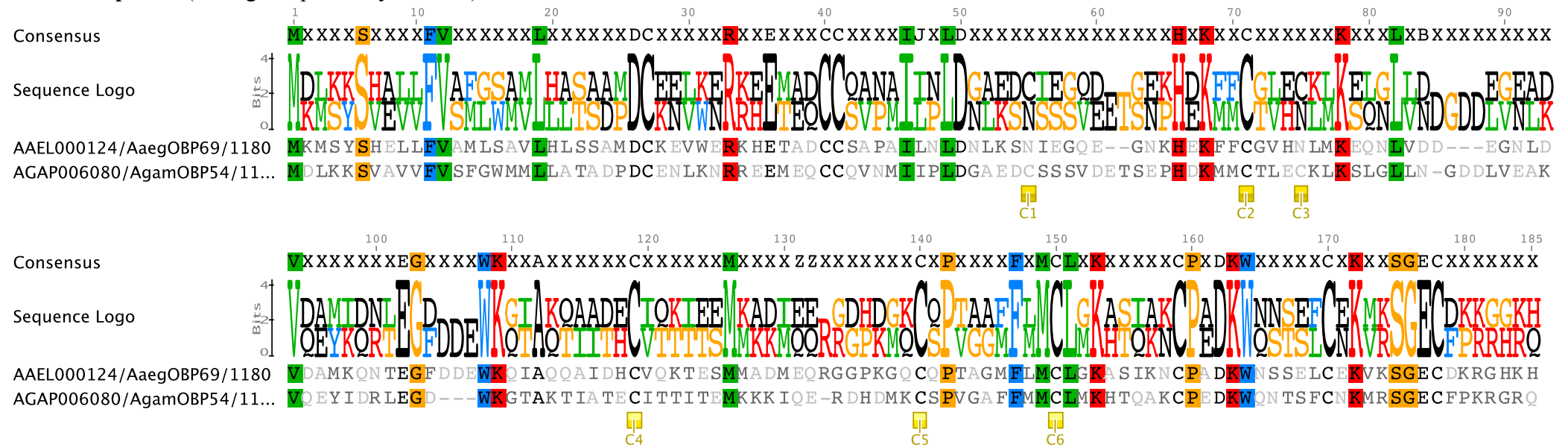
Plus C : *mplus4* (average seq. identity : 26.4%)Supplementary Figure 2b. Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

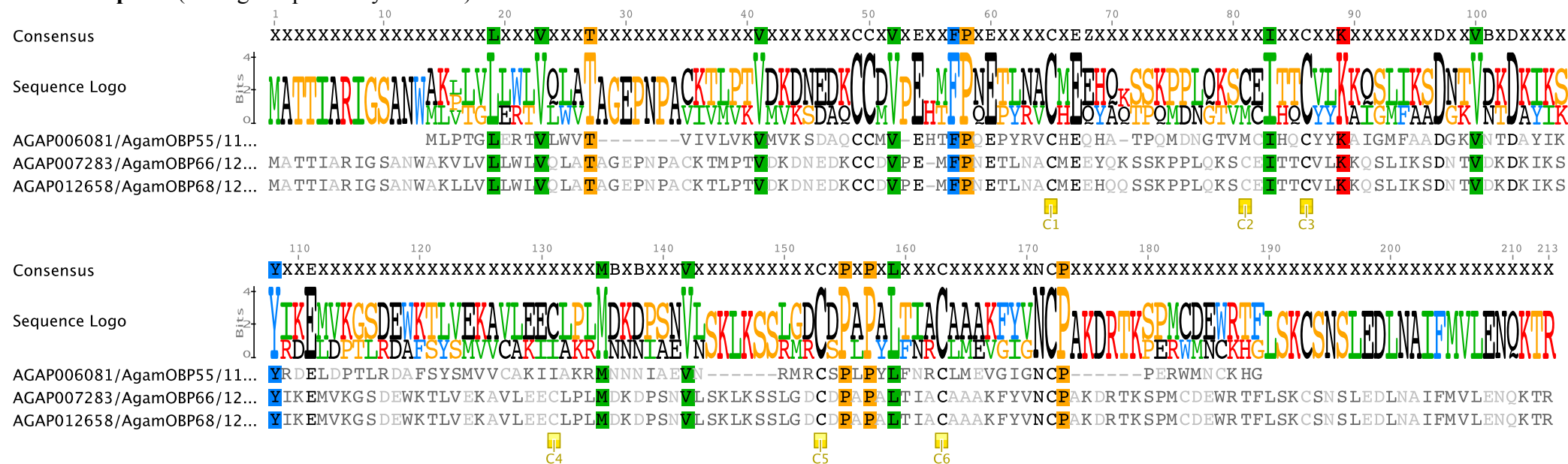
Plus C : mplus5 (average seq. identity : 56.4%)**Supplementary Figure 2b.** Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

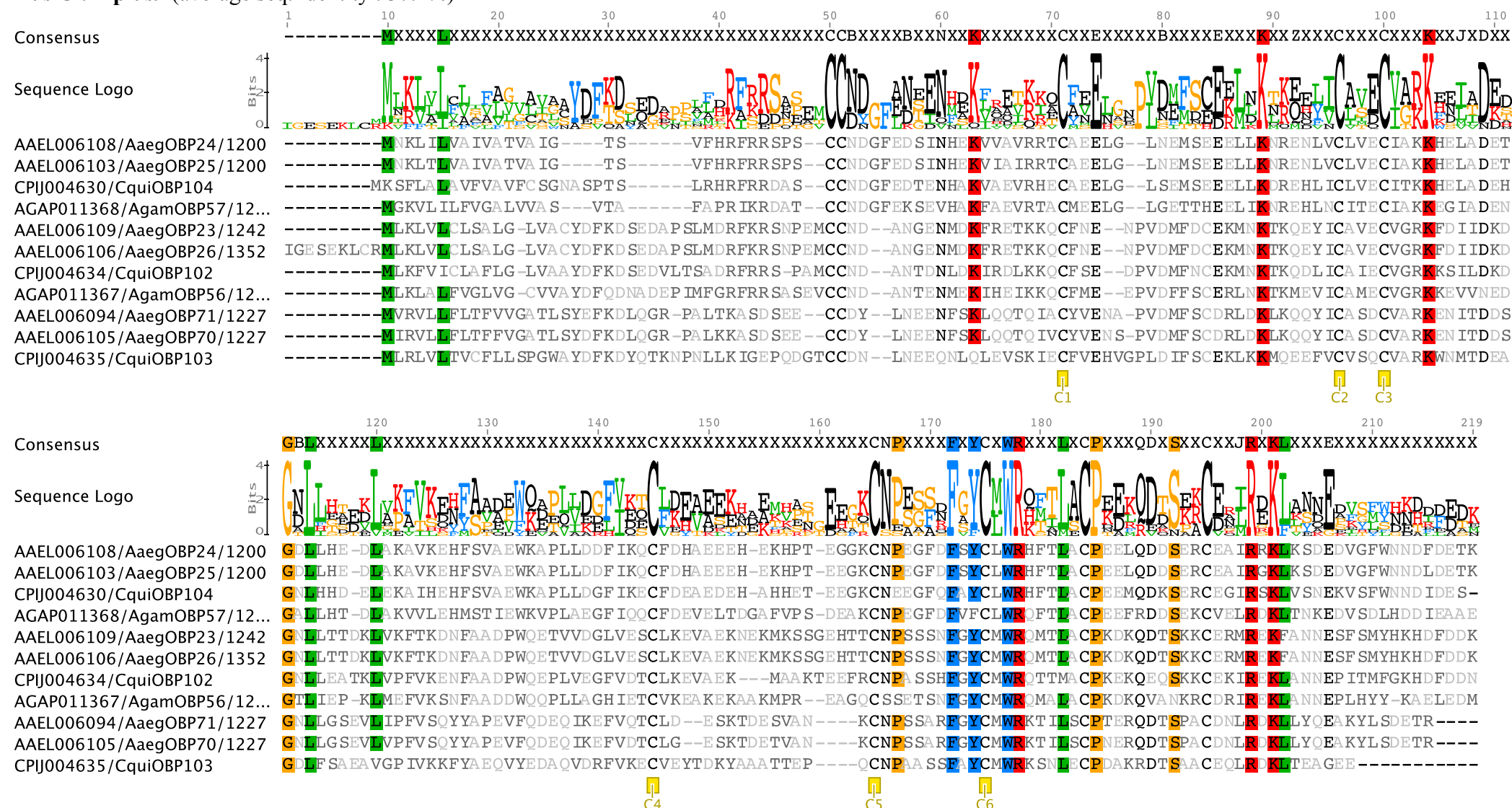
Plus C : mplus6 (average seq. identity : 37.1%)**Supplementary Figure 2b.** Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

Plus C : mplus7a (average seq. identity : 30.8%)**Supplementary Figure 2b.** Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

Plus C : mplus7b (average seq. identity : 29.3%)**Supplementary Figure 2b.** Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

Plus C : mplus7c (average seq. identity : 24.3%)

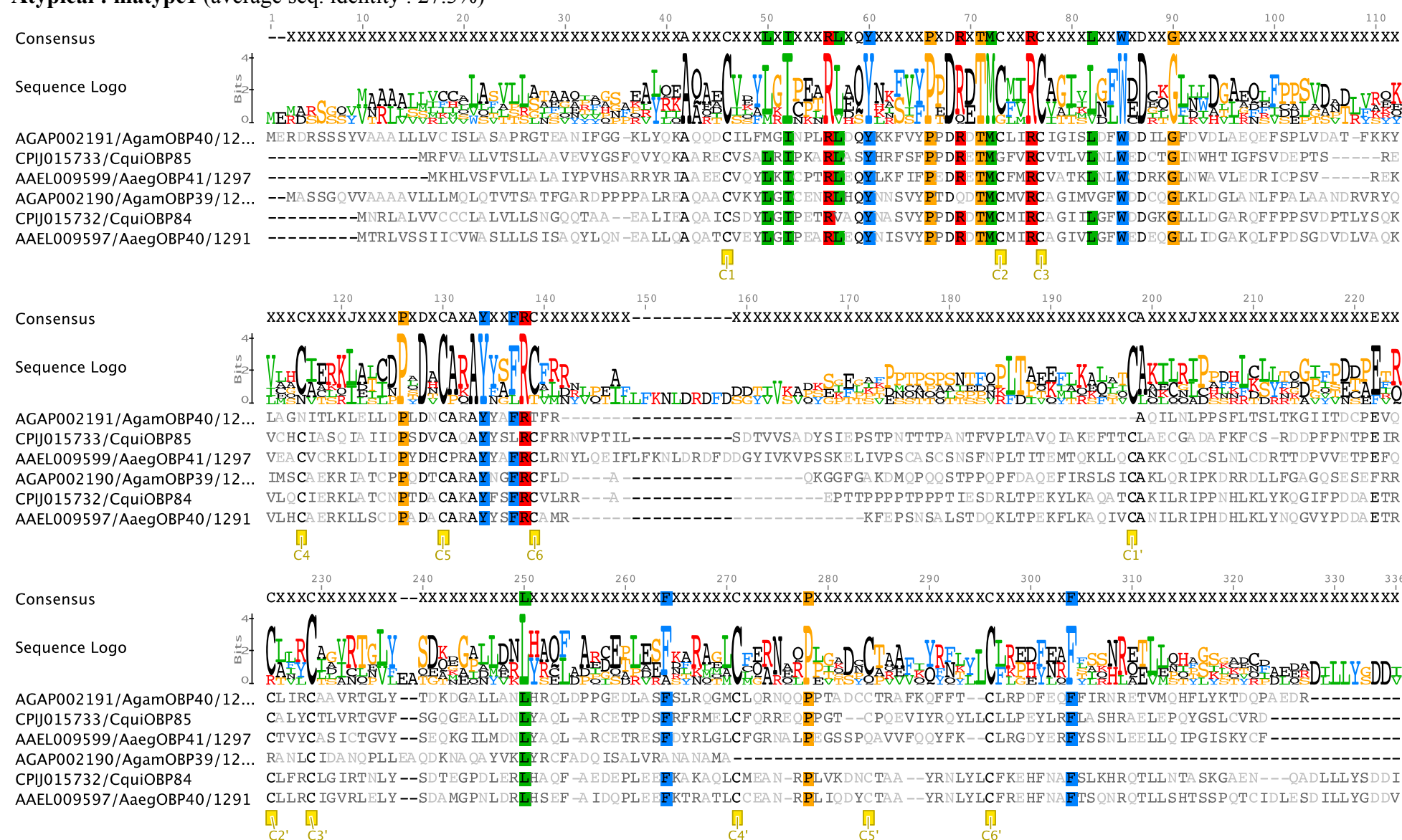
Plus C : mplus8 (average seq. identity : 48.2%)**Supplementary Figure 2b.** Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

Plus C : *mplus9* (average seq. identity : 37.2%)Supplementary Figure 2b. Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

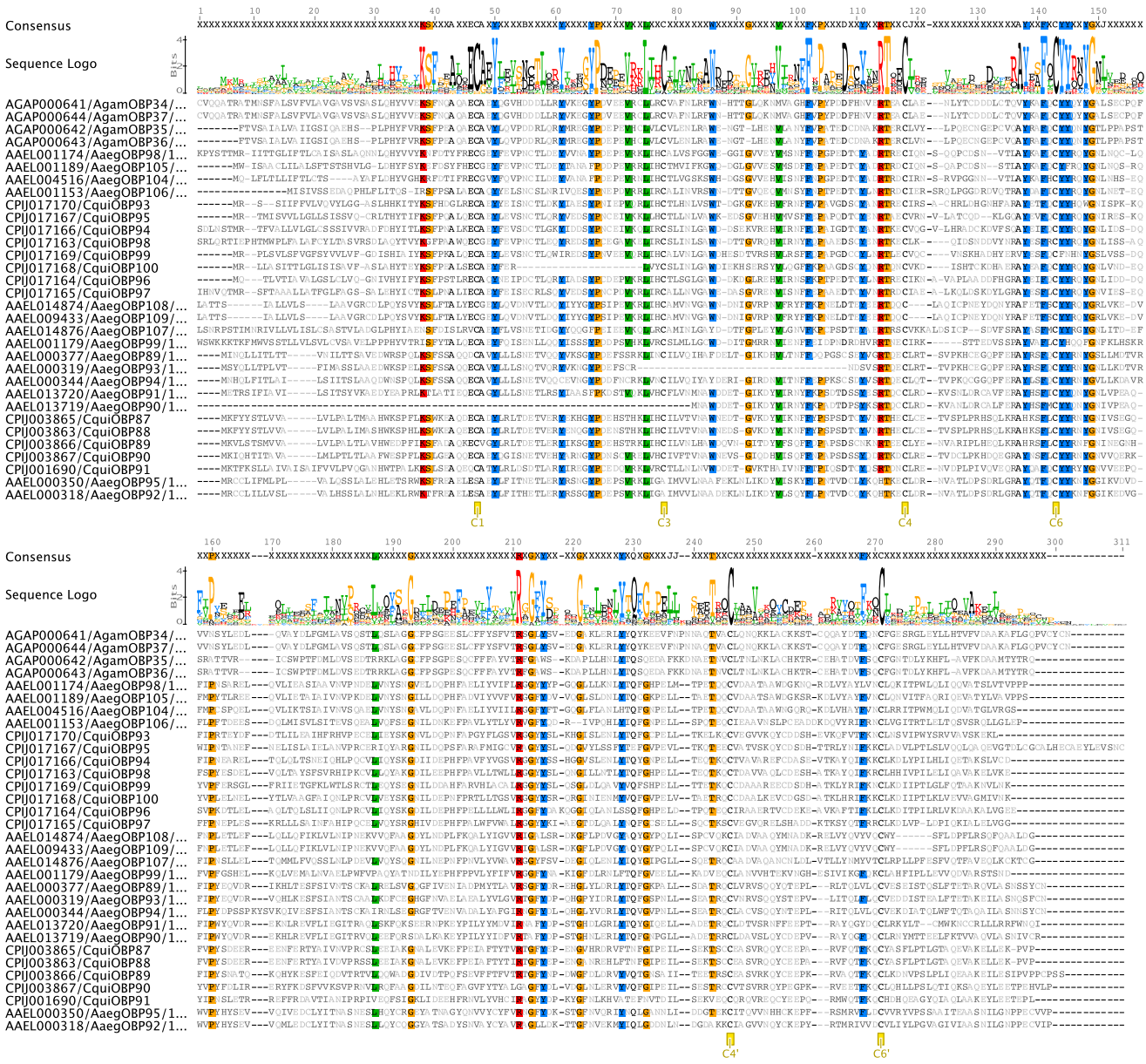
Additional legend to supplementary Figure 2b.

Shown are the alignments for the *PlusC* OBPs from the three mosquito genomes *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*. These were obtained after aligning the sequences with a structure-based profile using ClustalX software. The global *PlusC* OBP alignment was splitted into different parts corresponding to the phylogenetic clusters that is established in this study (see Figures 4 and 5). The residues that are highly conserved (with 75% or more degree of conservation) are highlighted in the sequences and in the sequence logos above the alignment. The consensus sequence is also featured on top of each alignment. The average pairwise sequence identities within each cluster are indicated. The six cysteines that are conserved between *PlusC* proteins and *Classic* OBPs are highlighted and denoted C1 to C6 in the alignments to ease the comparison between these two subfamilies. These diagrams were generated by the Geneious software.

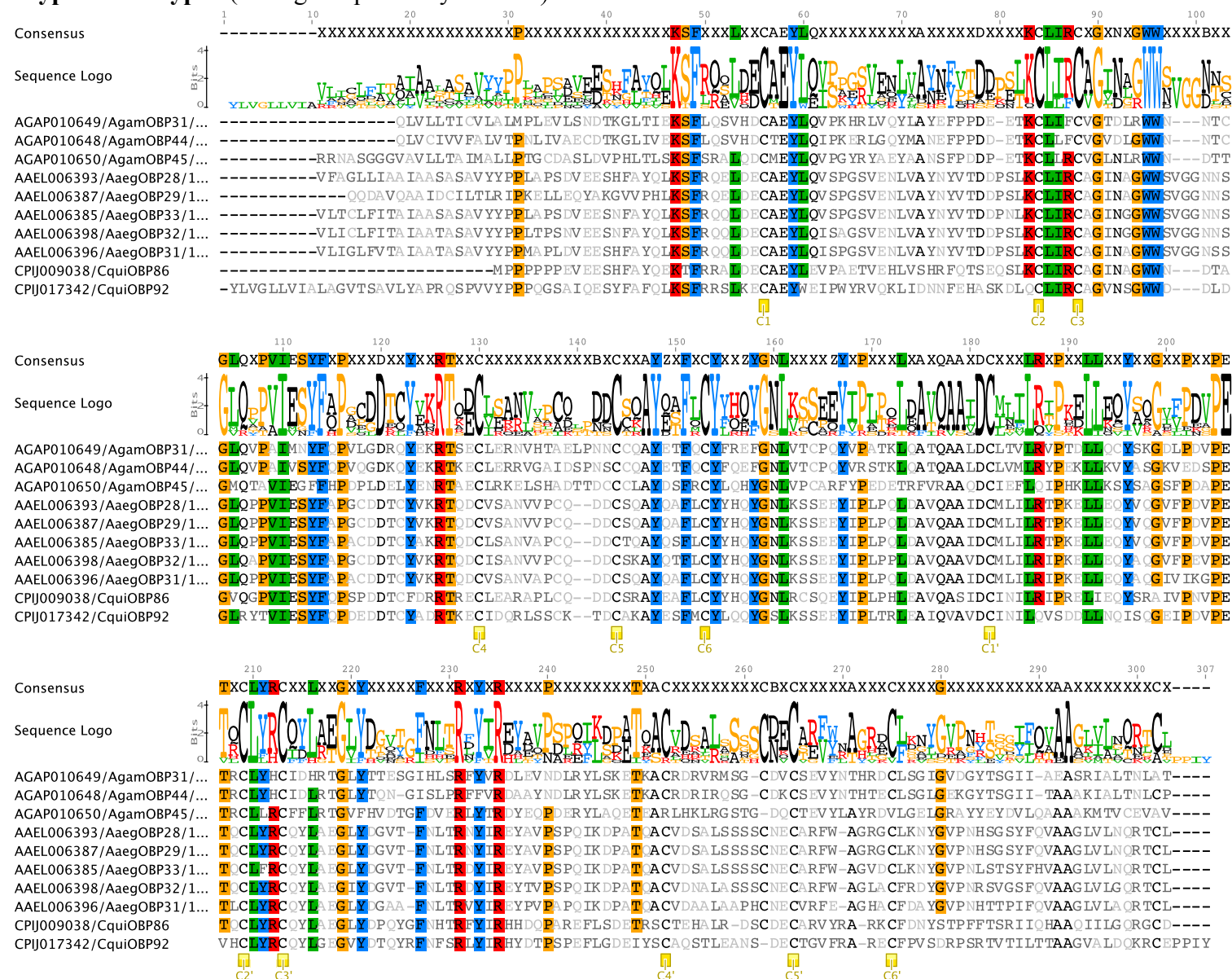
Supplementary Figure 2b. Protein sequence alignments of OBPs that belongs to the different clusters of the *PlusC* subfamily. See additional legend for details.

Atypical : matype1 (average seq. identity : 27.3%)**Supplementary Figure 2c.** Protein sequence alignments of OBPs that belongs to the different clusters of the *Atypical* subfamily. See additional legend for details.

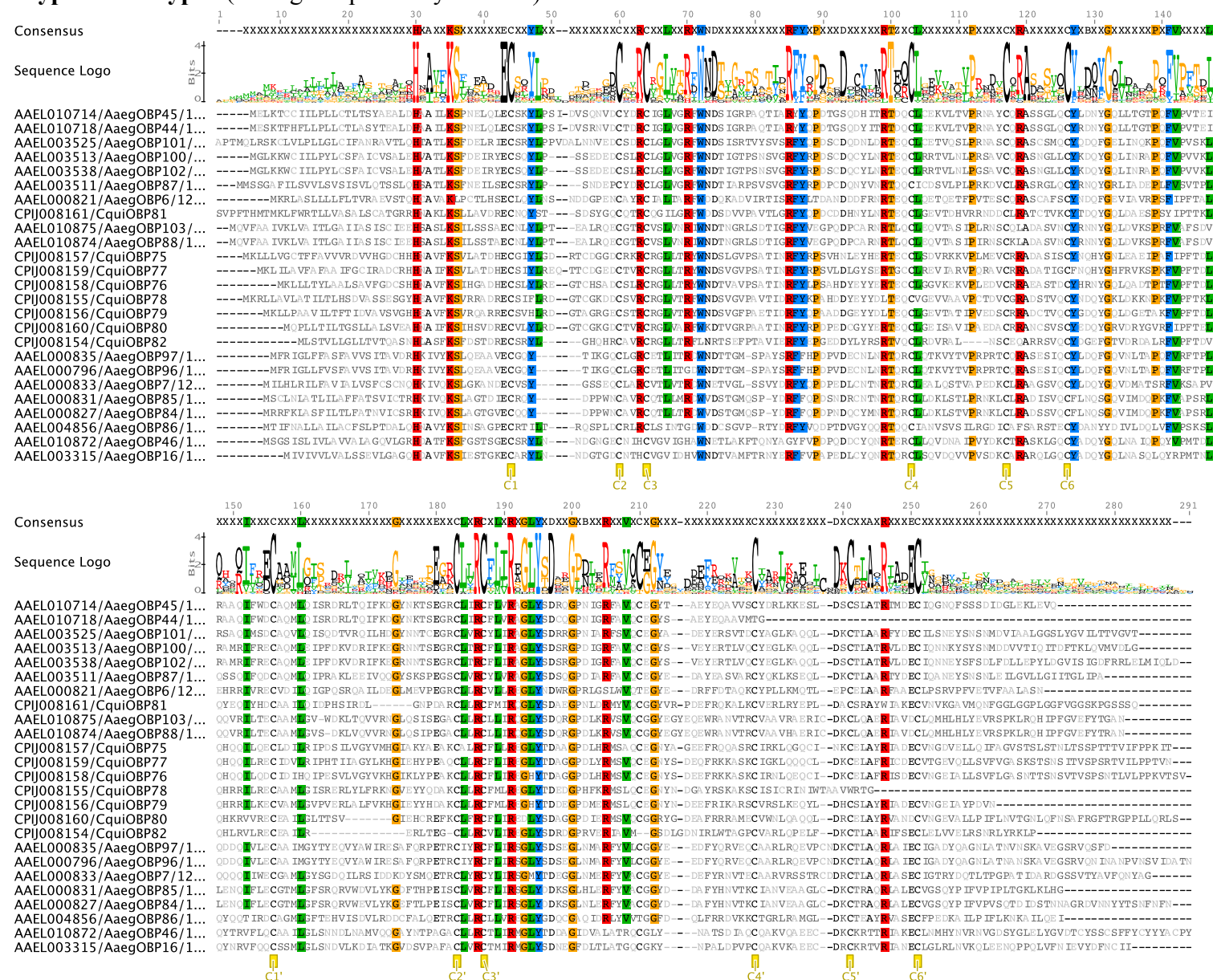
Atypical : matype2 (average seq. identity : 29.3%)



Supplementary Figure 2c. Protein sequence alignments of OBPs that belongs to the different clusters of the *Atypical* subfamily. See additional legend for details.

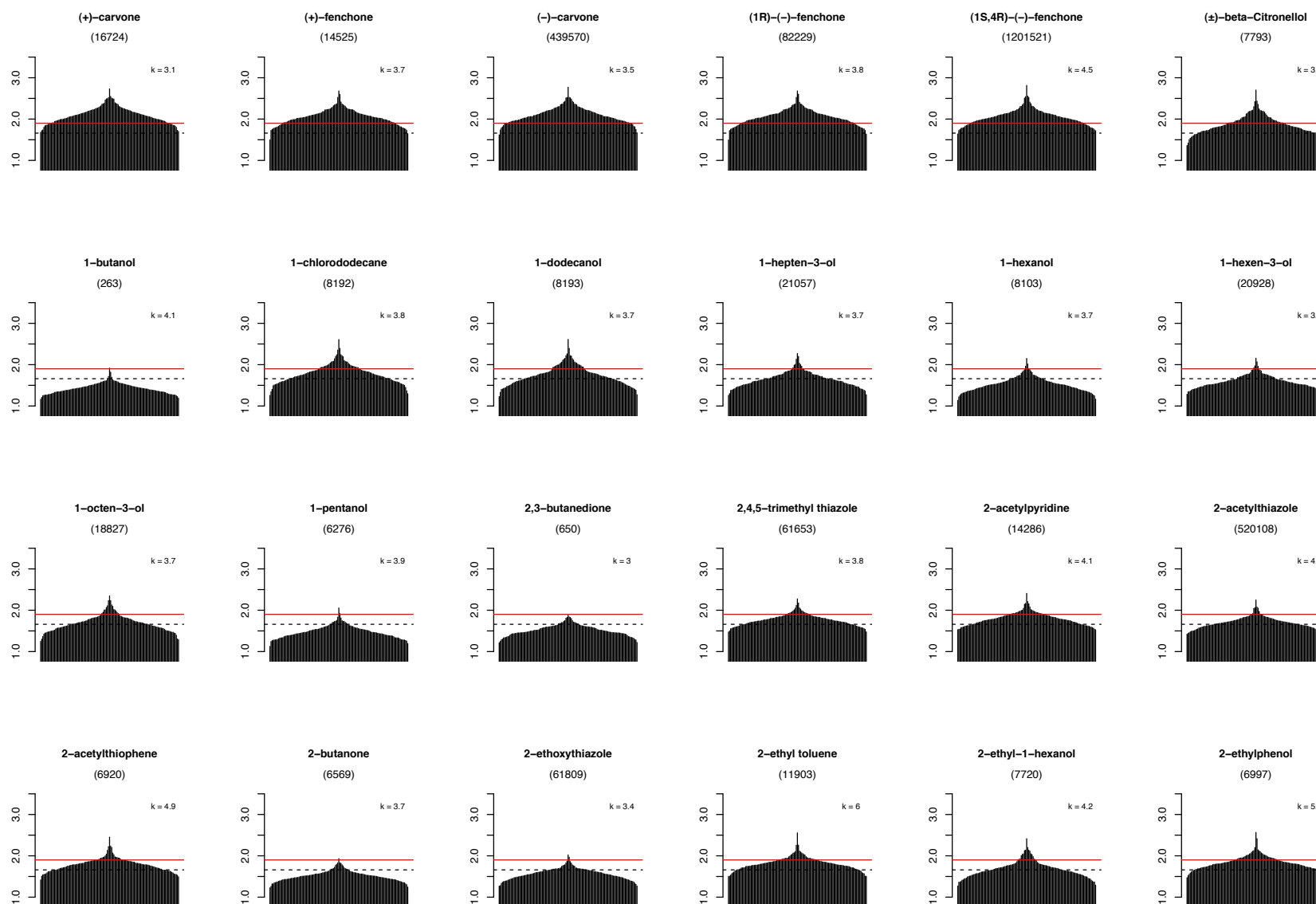
Atypical : matype3 (average seq. identity : 48.3%)**Supplementary Figure 2c.** Protein sequence alignments of OBPs that belongs to the different clusters of the *Atypical* subfamily. See additional legend for details.

Atypical : matype4 (average seq. identity : 33.5%)

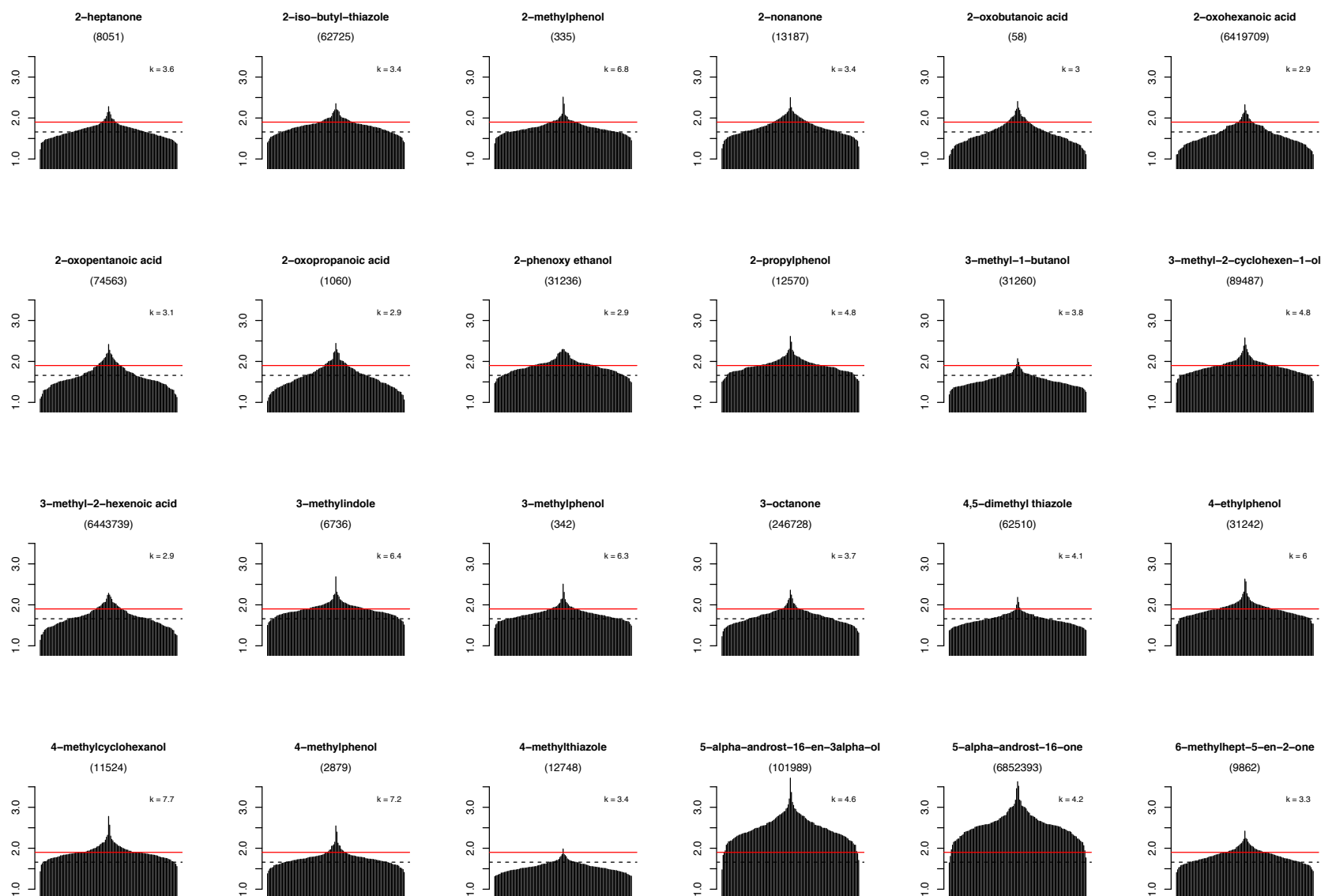
Supplementary Figure 2c. Protein sequence alignments of OBPs to the different clusters of the *Atypical* subfamily. See additional legend for details.

Additional legend to Figure 2c.

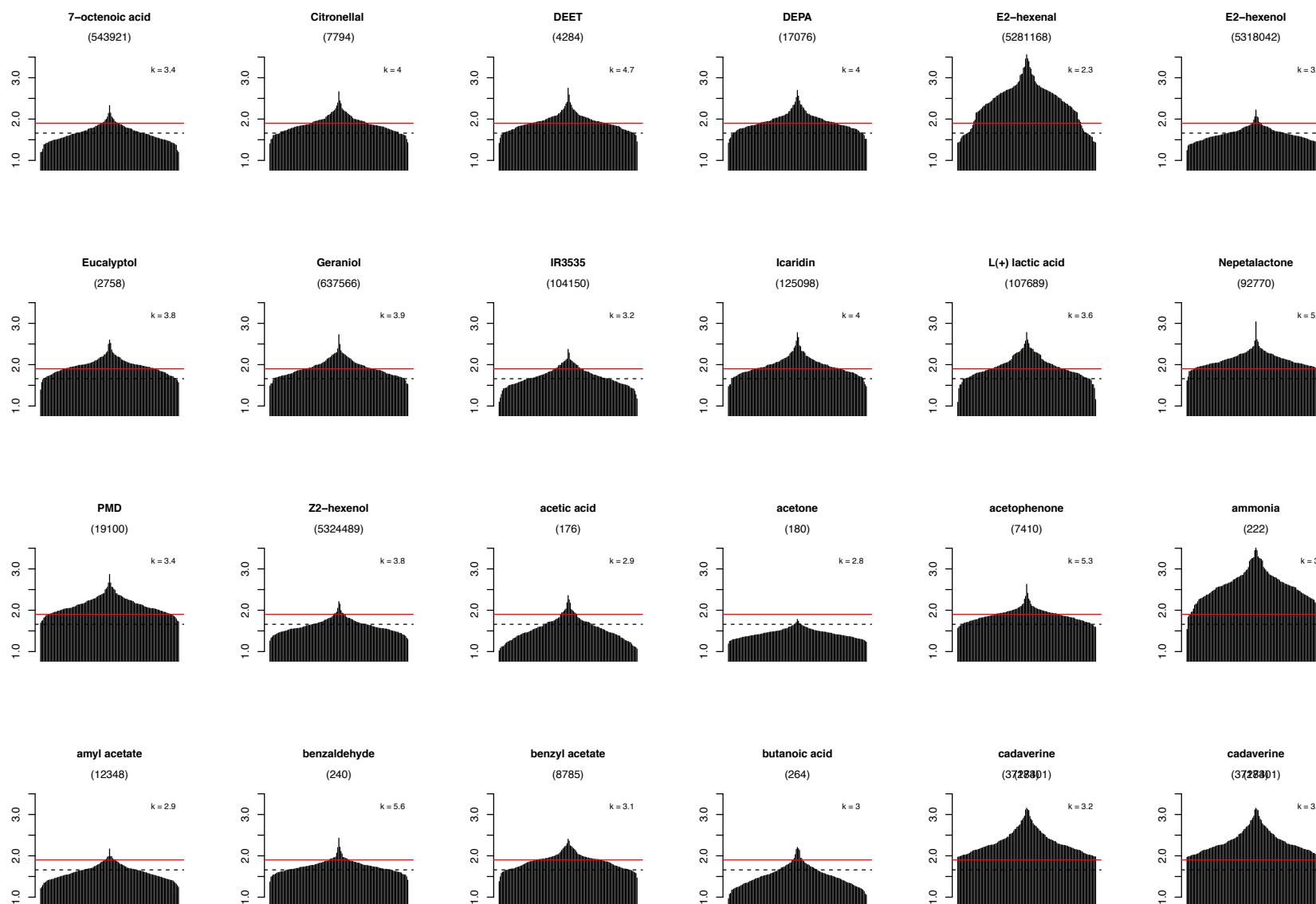
Shown are the alignments for the *Atypical* OBPs from the three mosquito genomes *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*. These were obtained after aligning the sequences with a structure-based profile using ClustalX software. The global *Atypical* OBP alignment was splitted into different parts corresponding to the phylogenetic clusters that is established in this study (see Figures 4 and 5). The residues that are highly conserved (with 75% or more degree of conservation) are highlighted in the sequences and in the sequence logos above the alignment. The consensus sequence is also featured on top of each alignment. The average pairwise sequence identities within each cluster are indicated. The six cysteines that are conserved between the two constitutive OBP domains of *Atypical* proteins and *Classic* OBPs are highlighted and denoted C1 to C6 in the N-term domain and C1' to C6' in the C-term domain in the alignments to ease the comparison between these two subfamilies. These diagrams were generated by the Geneious software.



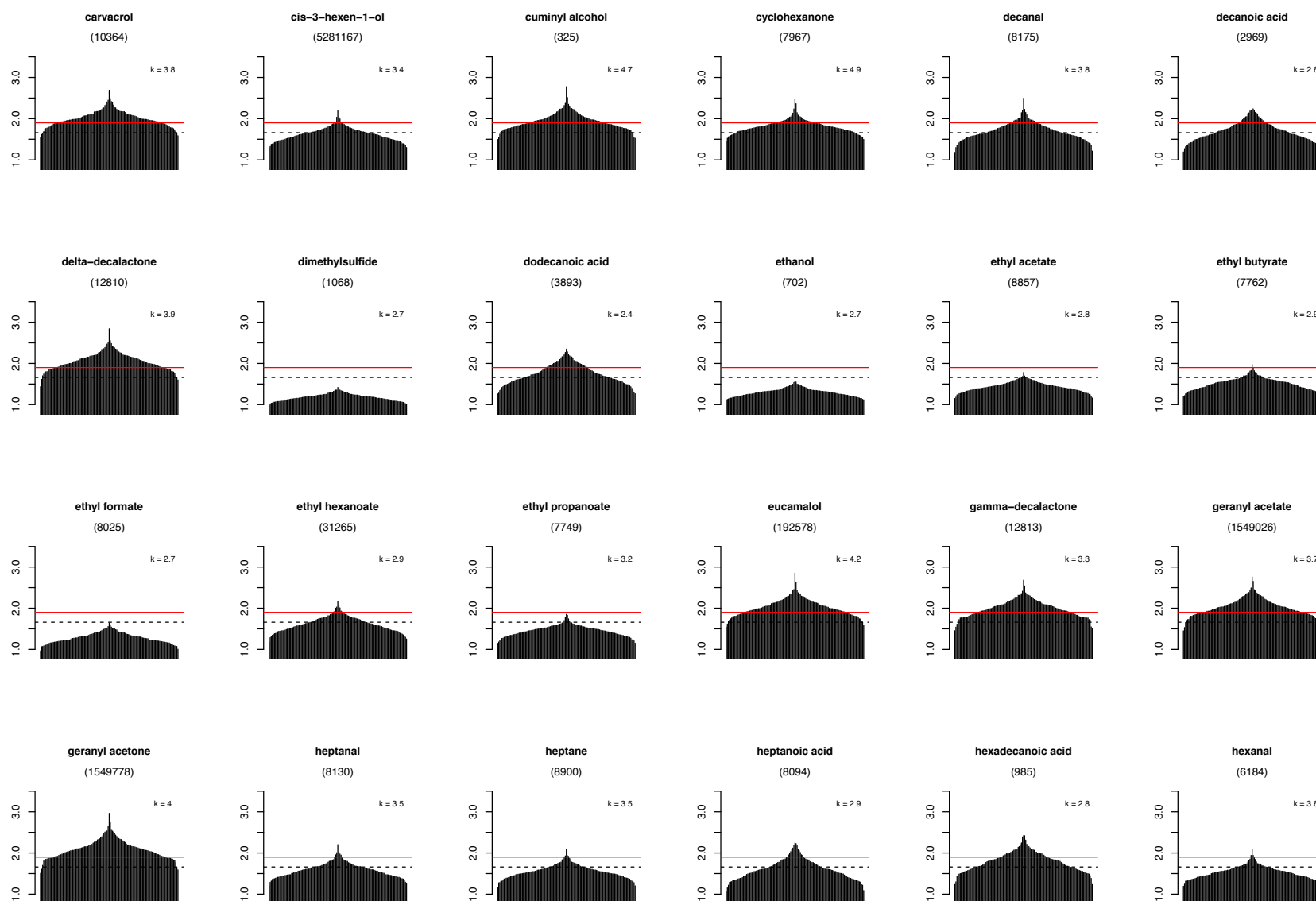
Supplementary Figure 3a. Ligand tuning curves rendered as a measure of specificity of the ligand towards the odorant binding proteins in the mosquito genome. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking proteins are centered in the middle of the plots and low ranking proteins towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.



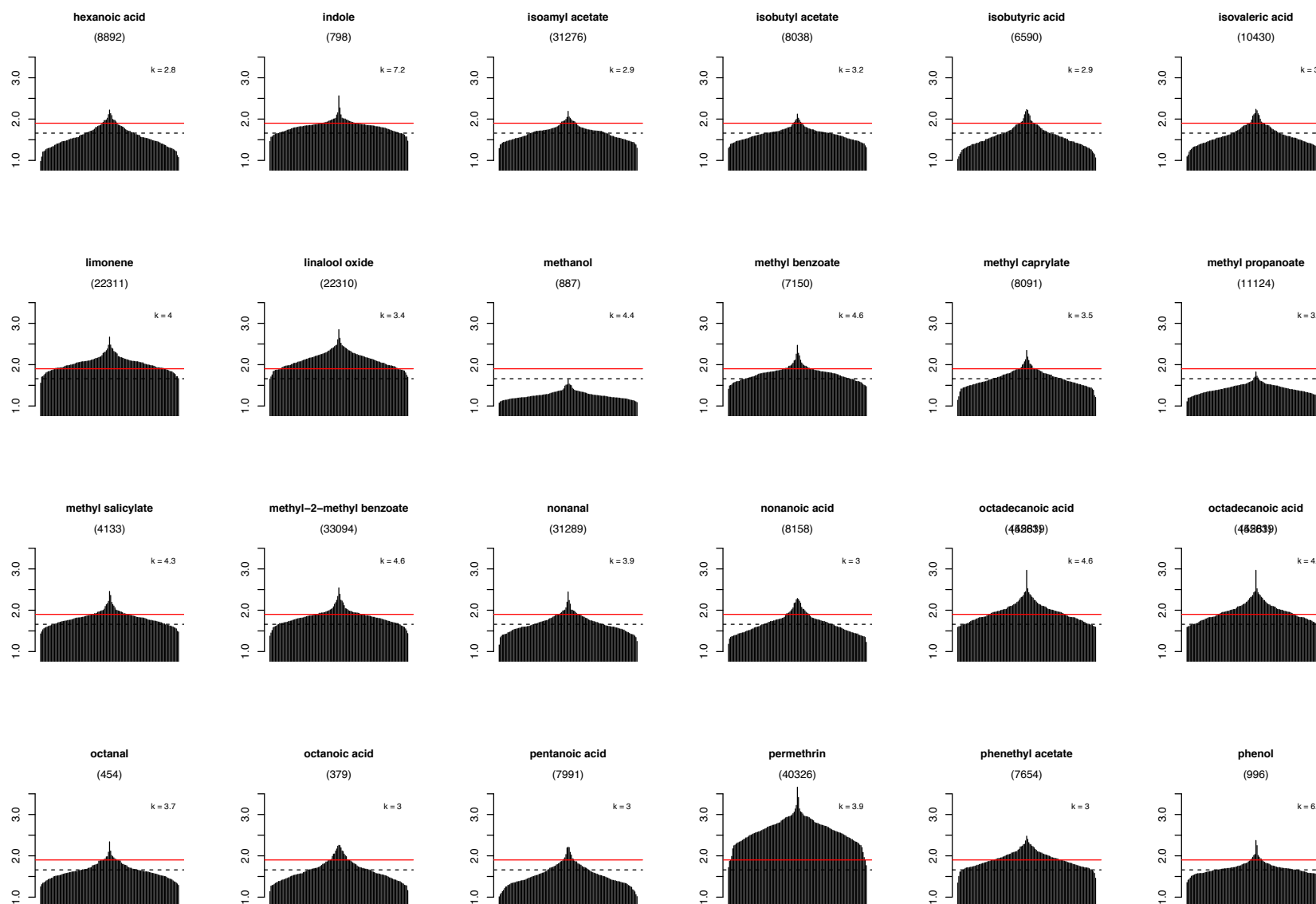
Supplementary Figure 3a. Ligand tuning curves rendered as a measure of specificity of the ligand towards the odorant binding proteins in the mosquito genome. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking proteins are centered in the middle of the plots and low ranking proteins towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.



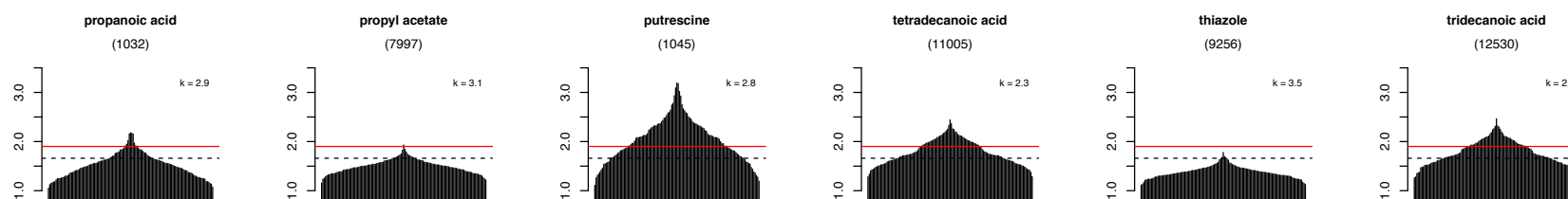
Supplementary Figure 3a. Ligand tuning curves rendered as a measure of specificity of the ligand towards the odorant binding proteins in the mosquito genome. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking proteins are centered in the middle of the plots and low ranking proteins towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.



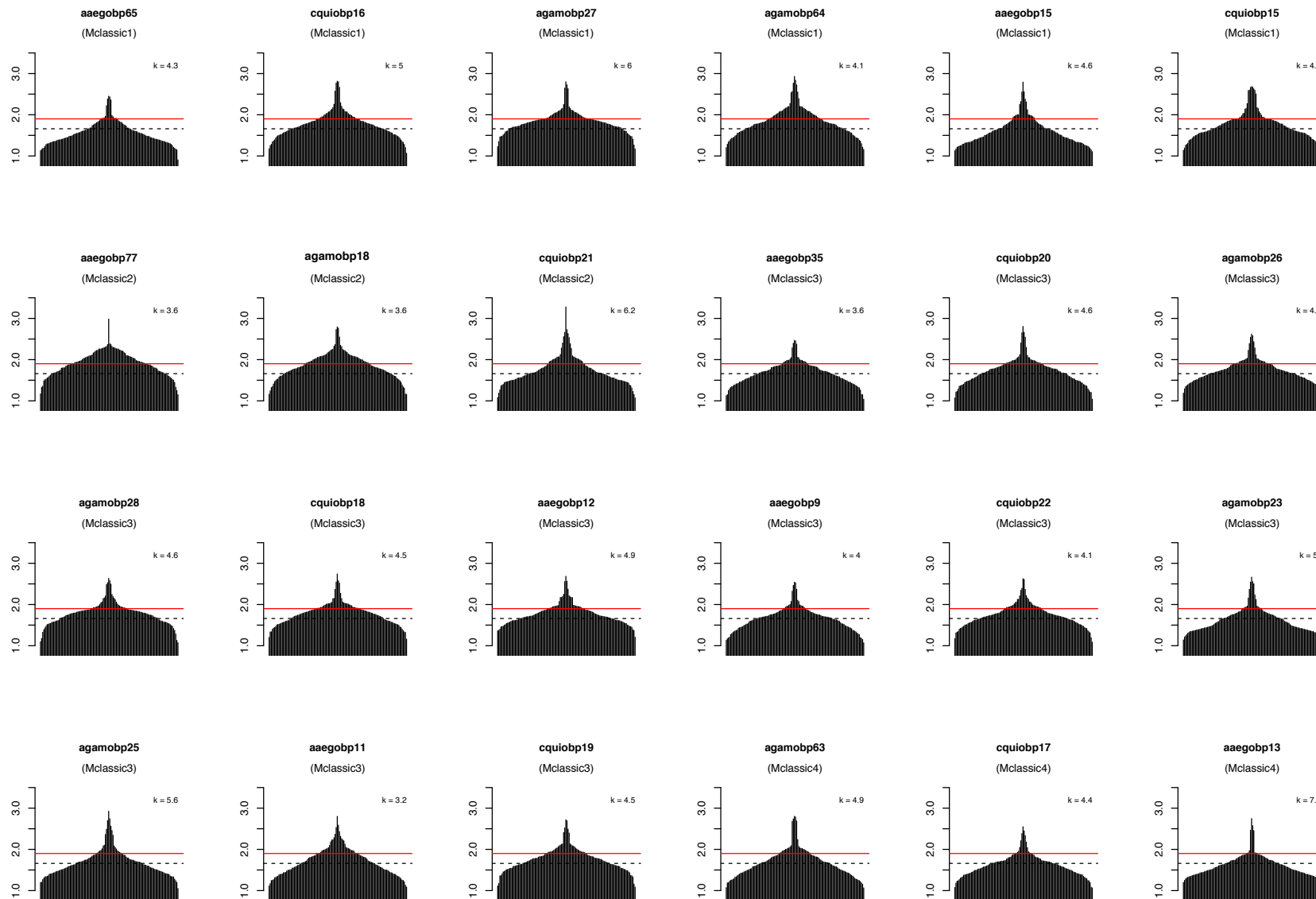
Supplementary Figure 3a. Ligand tuning curves rendered as a measure of specificity of the ligand towards the odorant binding proteins in the mosquito genome. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking proteins are centered in the middle of the plots and low ranking proteins towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.



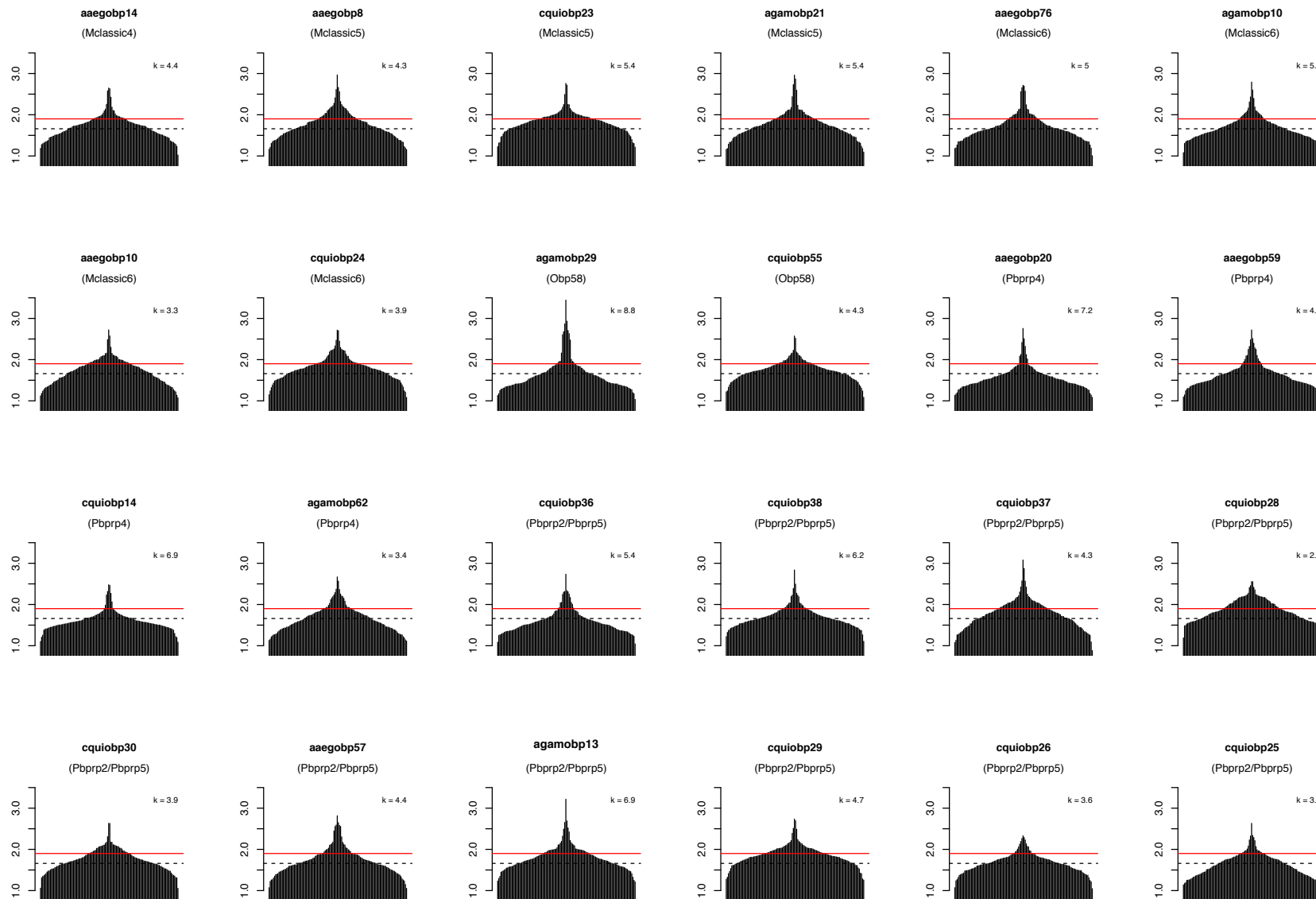
Supplementary Figure 3a. Ligand tuning curves rendered as a measure of specificity of the ligand towards the odorant binding proteins in the mosquito genome. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking proteins are centered in the middle of the plots and low ranking proteins towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.



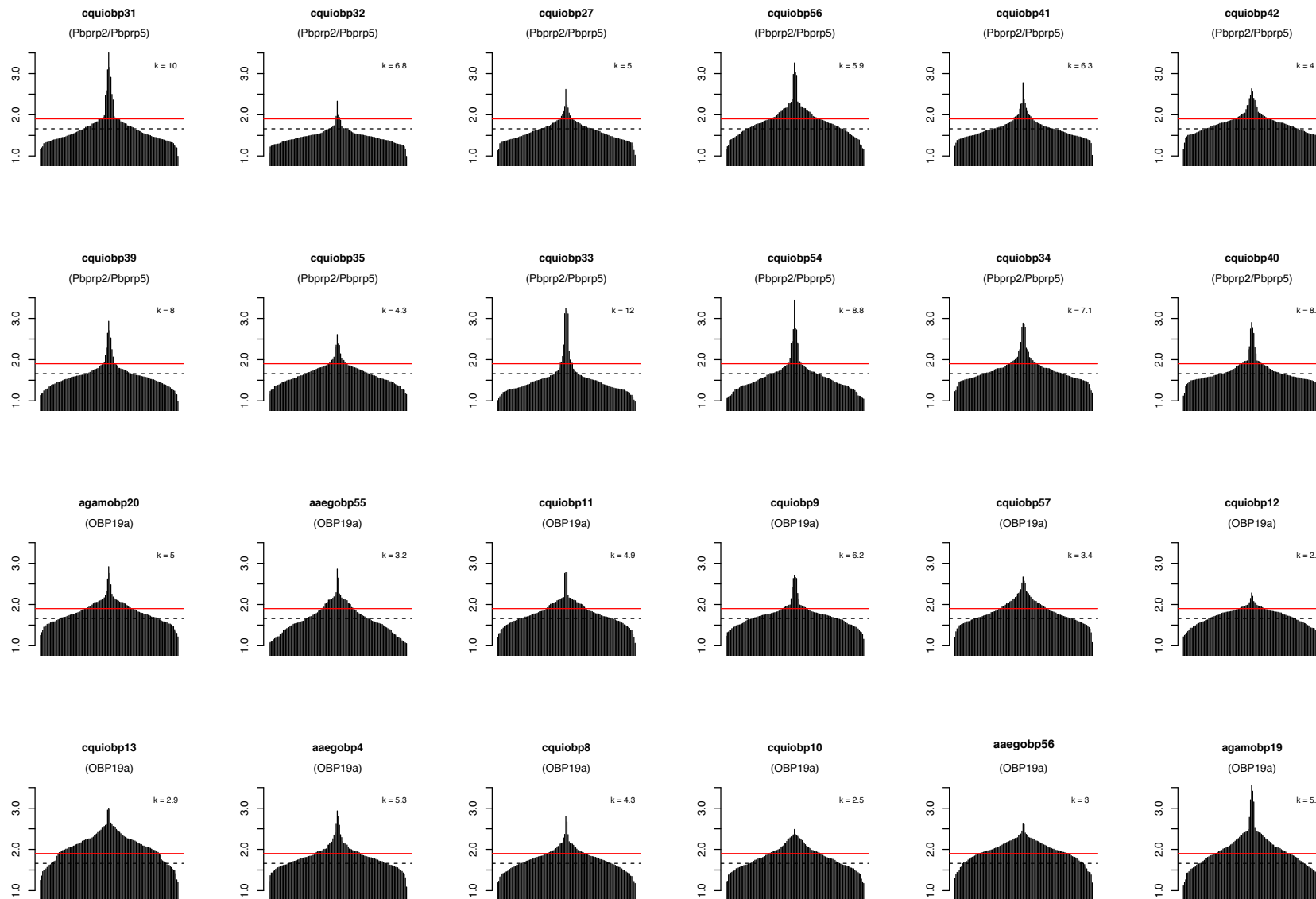
Supplementary Figure 3a. Ligand tuning curves rendered as a measure of specificity of the ligand towards the odorant binding proteins in the mosquito genome. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking proteins are centered in the middle of the plots and low ranking proteins towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that ligand based on the SILE values.



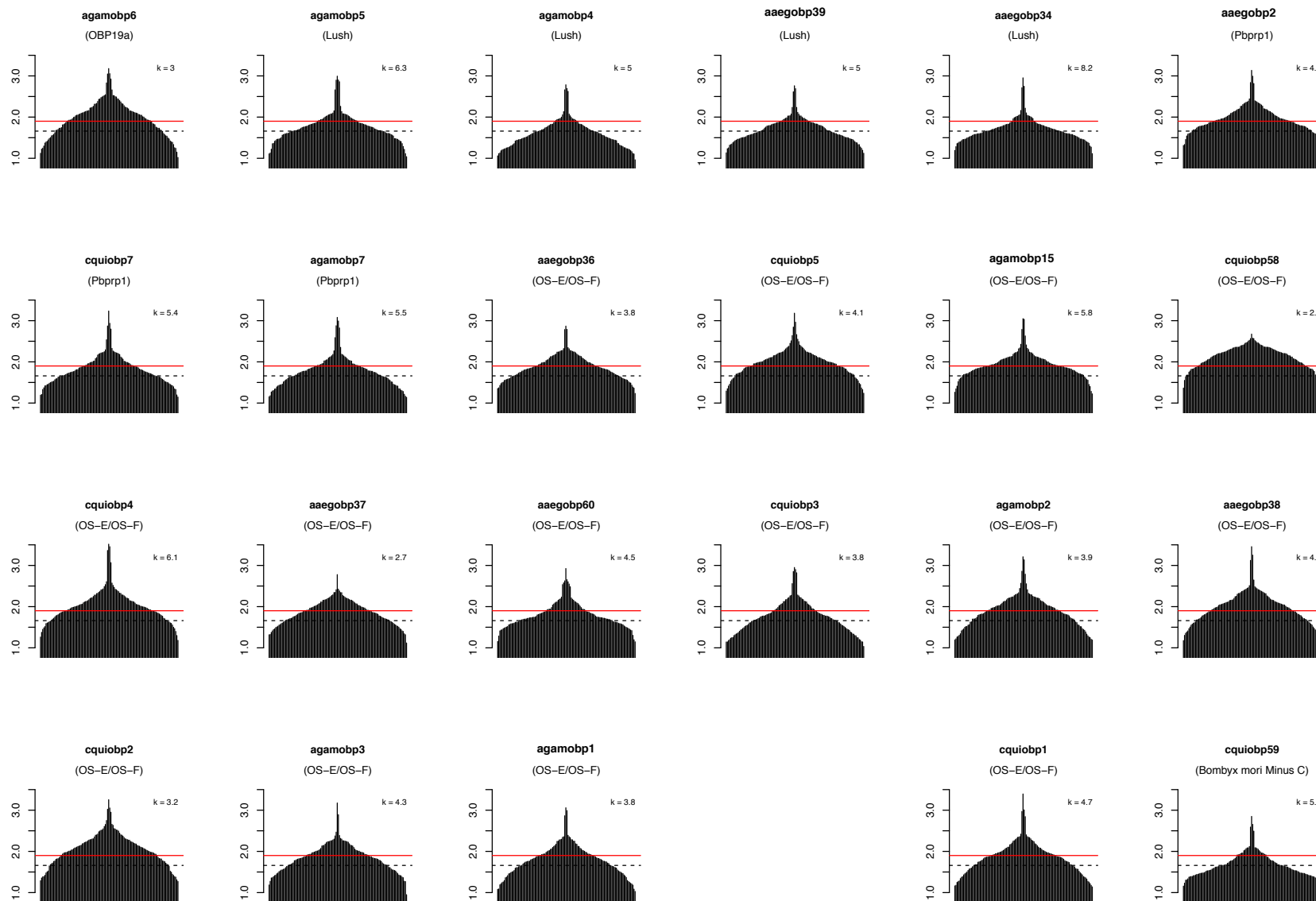
Supplementary Figure 3b. OBP tuning curves rendered as a measure of specificity of a given OBP towards a given set of odorants in the mosquito genome.. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking odorants are centered in the middle of the plots and low ranking ligands towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that OBP based on the SILE values.



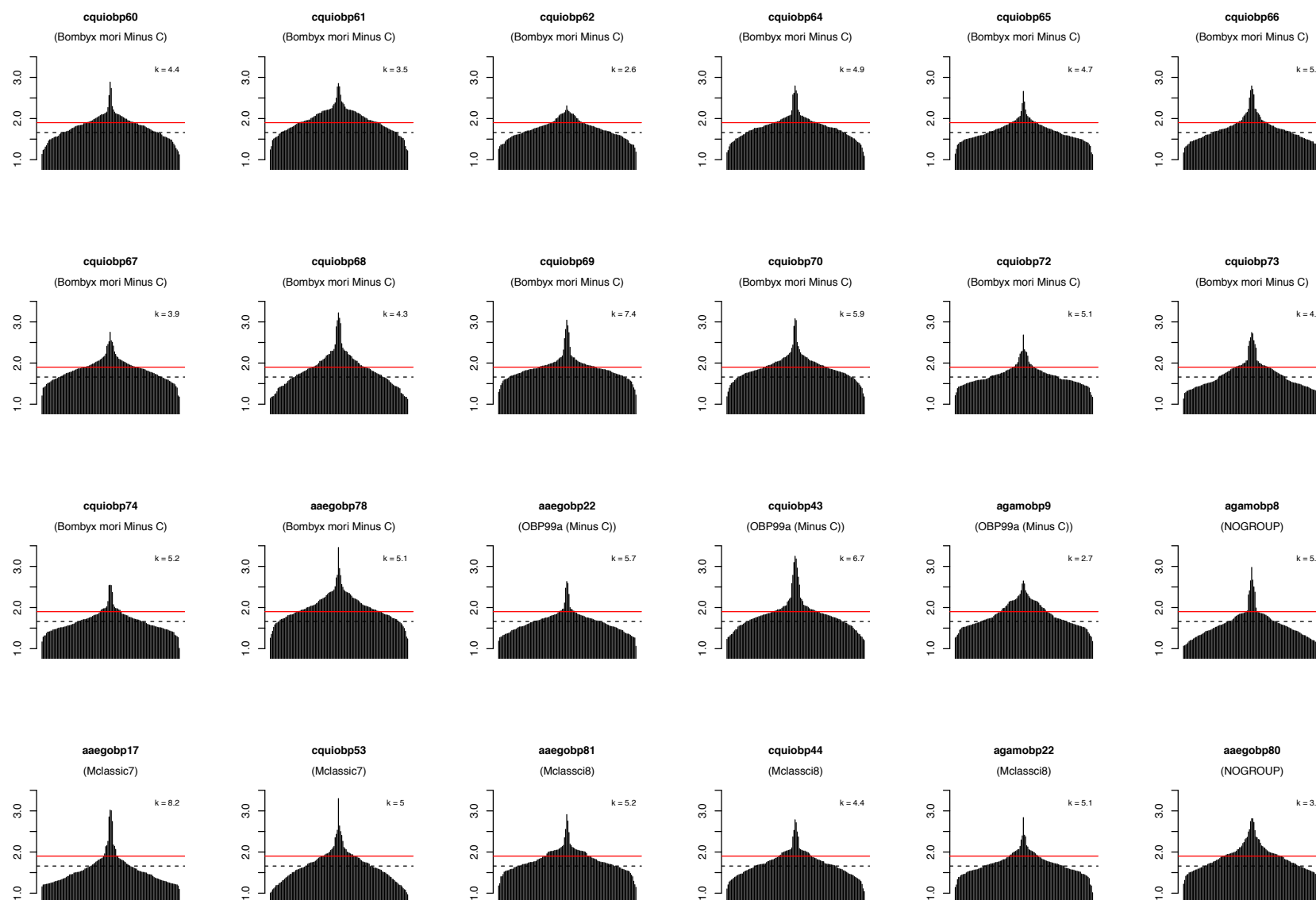
Supplementary Figure 3b. OBP tuning curves rendered as a measure of specificity of a given OBP towards a given set of odorants in the mosquito genome.. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking odorants are centered in the middle of the plots and low ranking ligands towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that OBP based on the SILE values.



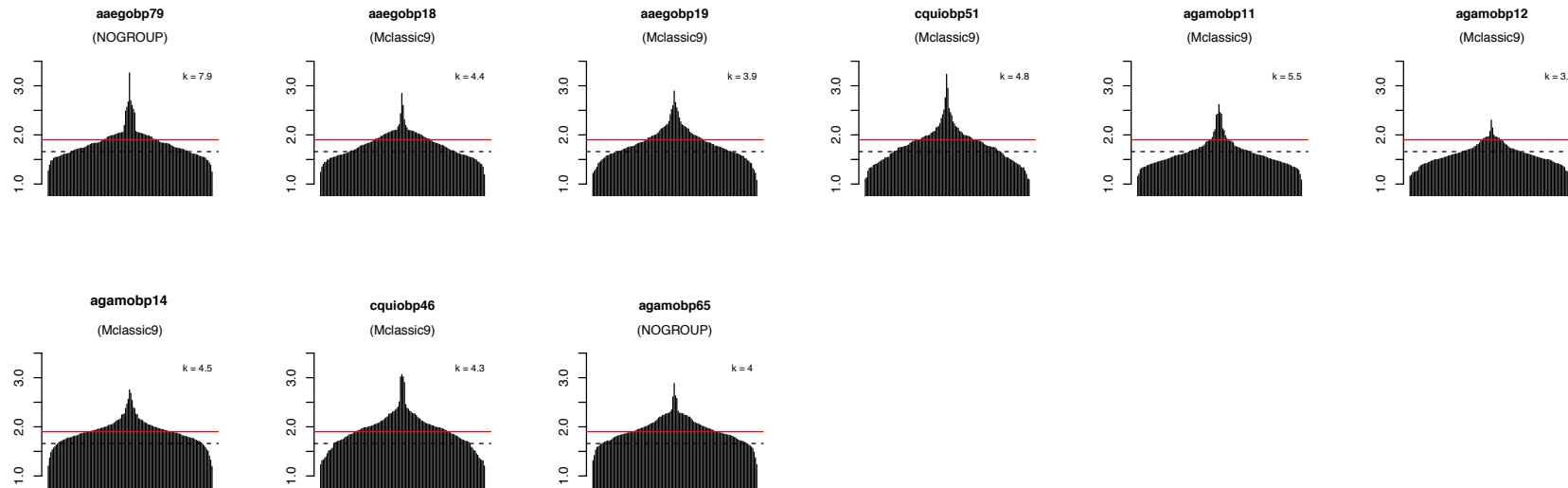
Supplementary Figure 3b. OBP tuning curves rendered as a measure of specificity of a given OBP towards a given set of odorants in the mosquito genome.. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking odorants are centered in the middle of the plots and low ranking ligands towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that OBP based on the SILE values.



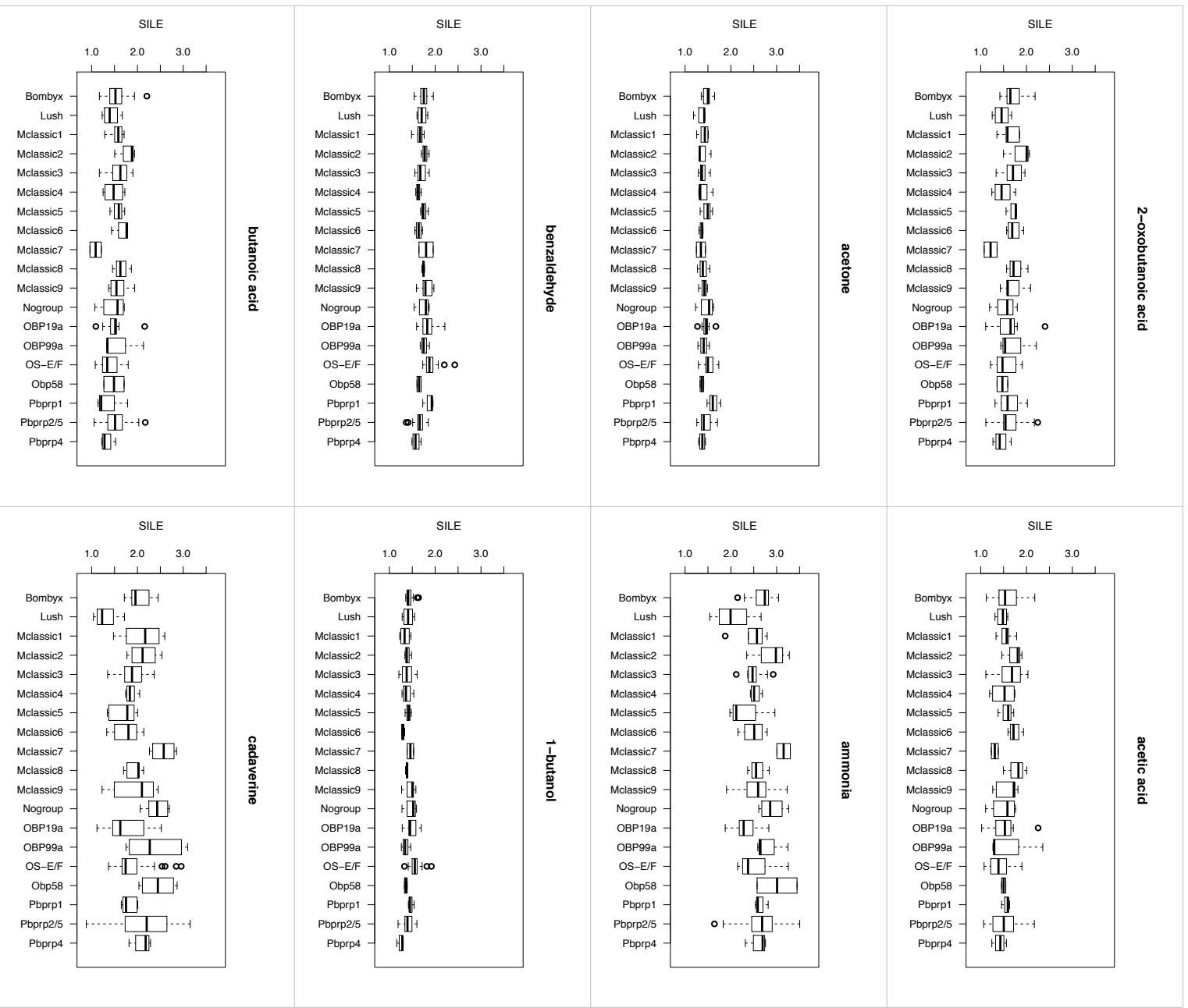
Supplementary Figure 3b. OBP tuning curves rendered as a measure of specificity of a given OBP towards a given set of odorants in the mosquito genome.. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking odorants are centered in the middle of the plots and low ranking ligands towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that OBP based on the SILE values.



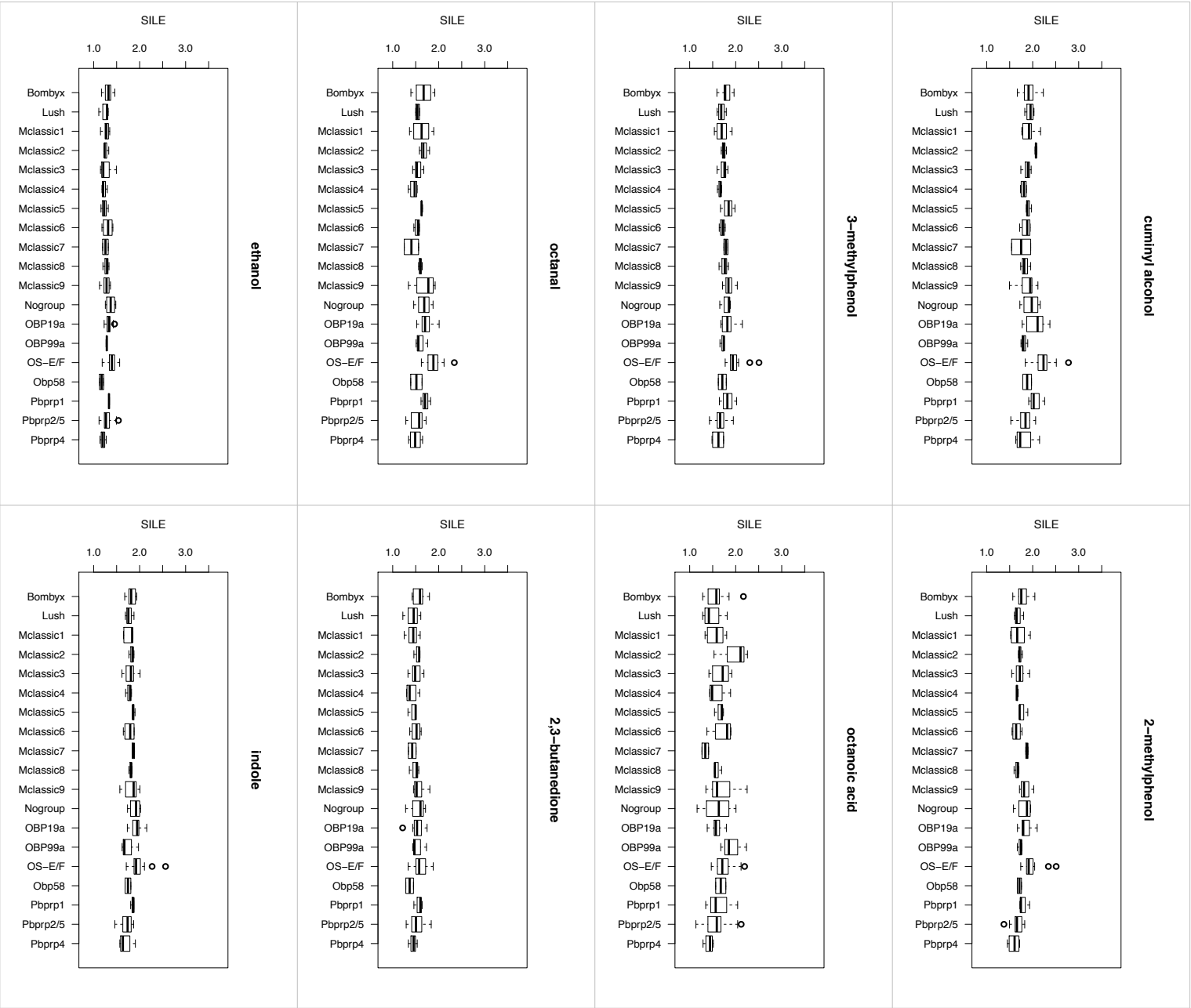
Supplementary Figure 3b. OBP tuning curves rendered as a measure of specificity of a given OBP towards a given set of odorants in the mosquito genome.. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking odorants are centered in the middle of the plots and low ranking ligands towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that OBP based on the SILE values.



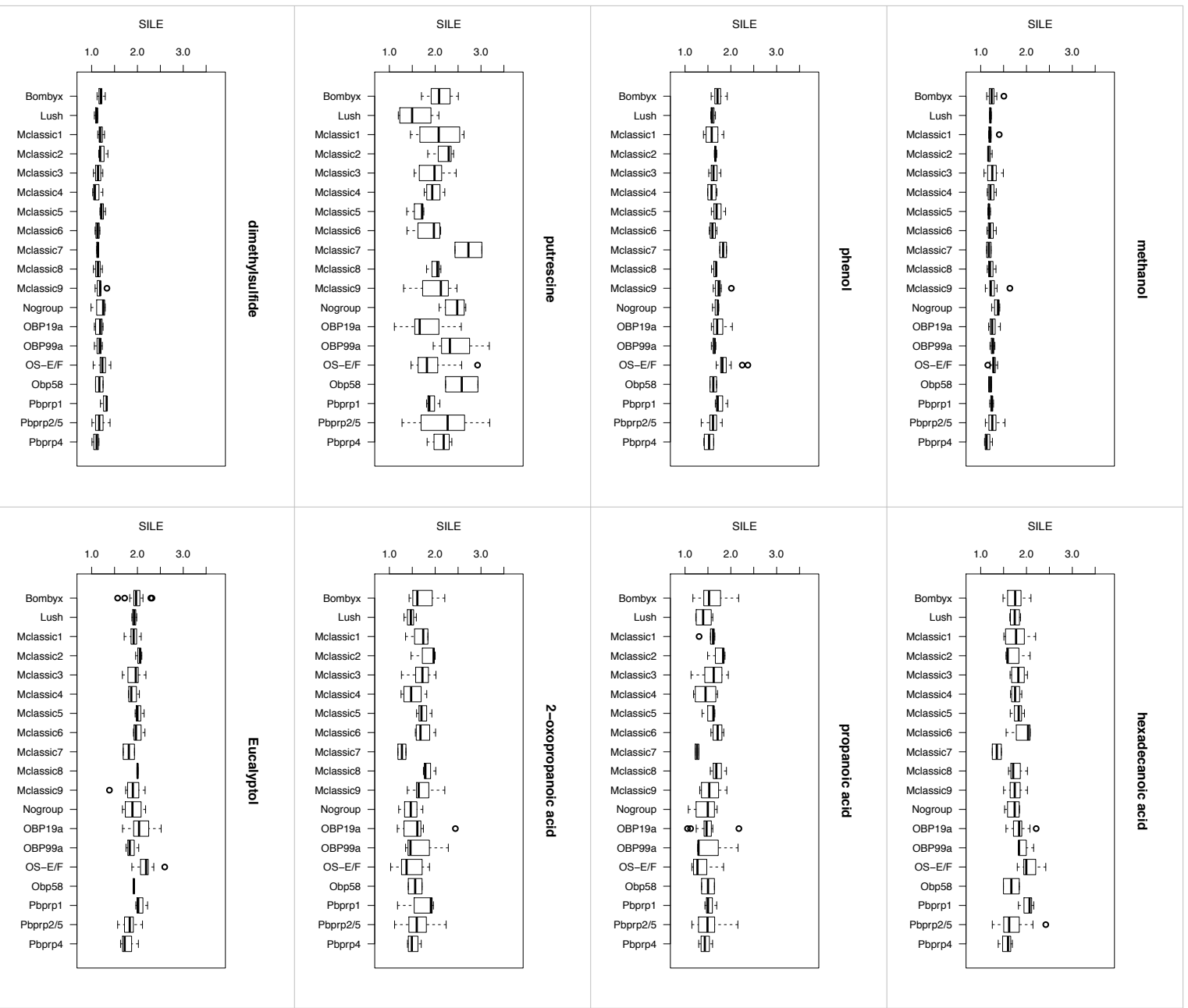
Supplementary Figure 3b. OBP tuning curves rendered as a measure of specificity of a given OBP towards a given set of odorants in the mosquito genome.. The size independent ligand efficiency (SILE) values are defined on the Y-axis and the OBPs along the X-axis. Top ranking odorants are centered in the middle of the plots and low ranking ligands towards the edges of the plots. The variable k in the plot represents the kurtosis value calculated for that OBP based on the SILE values.



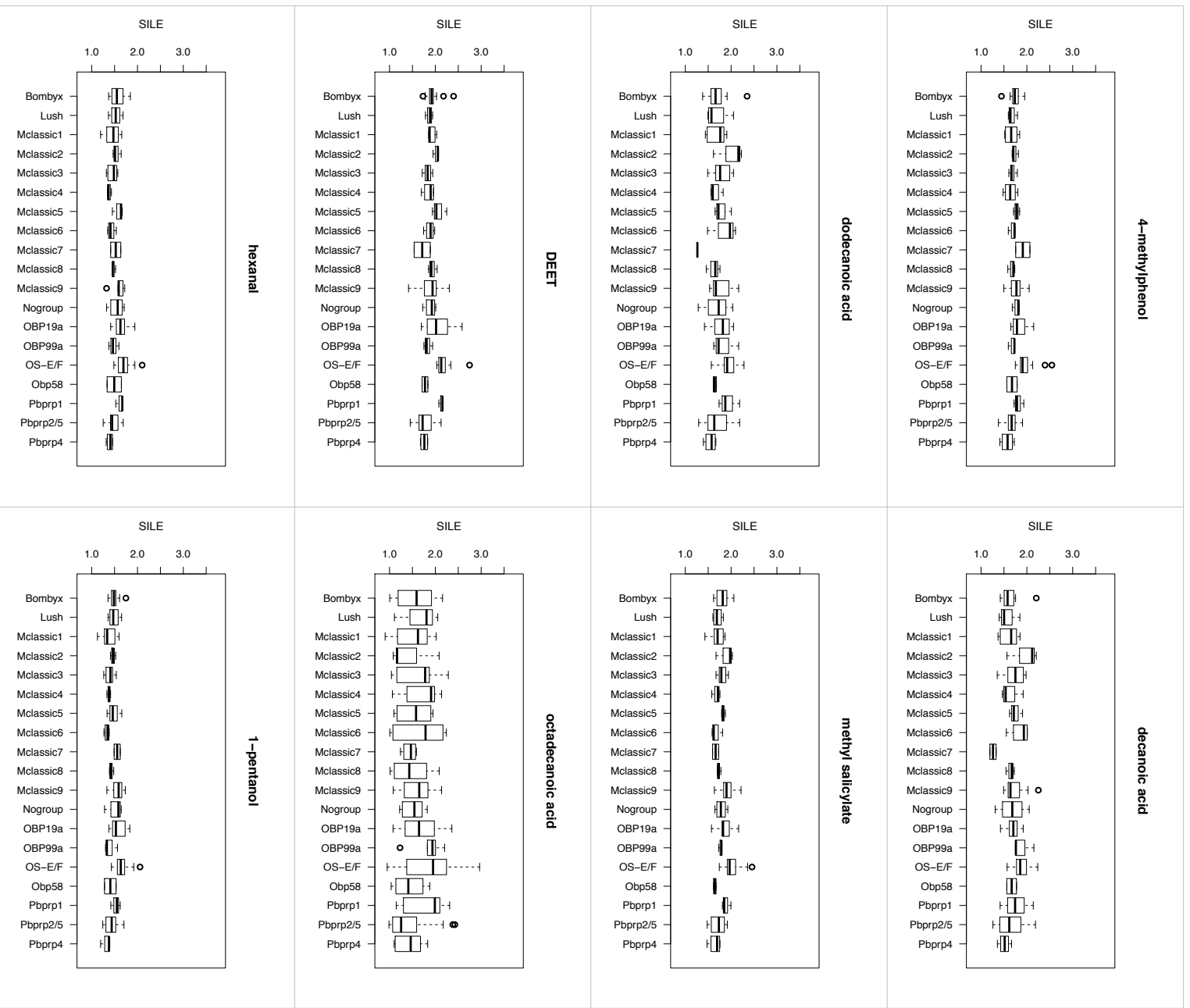
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



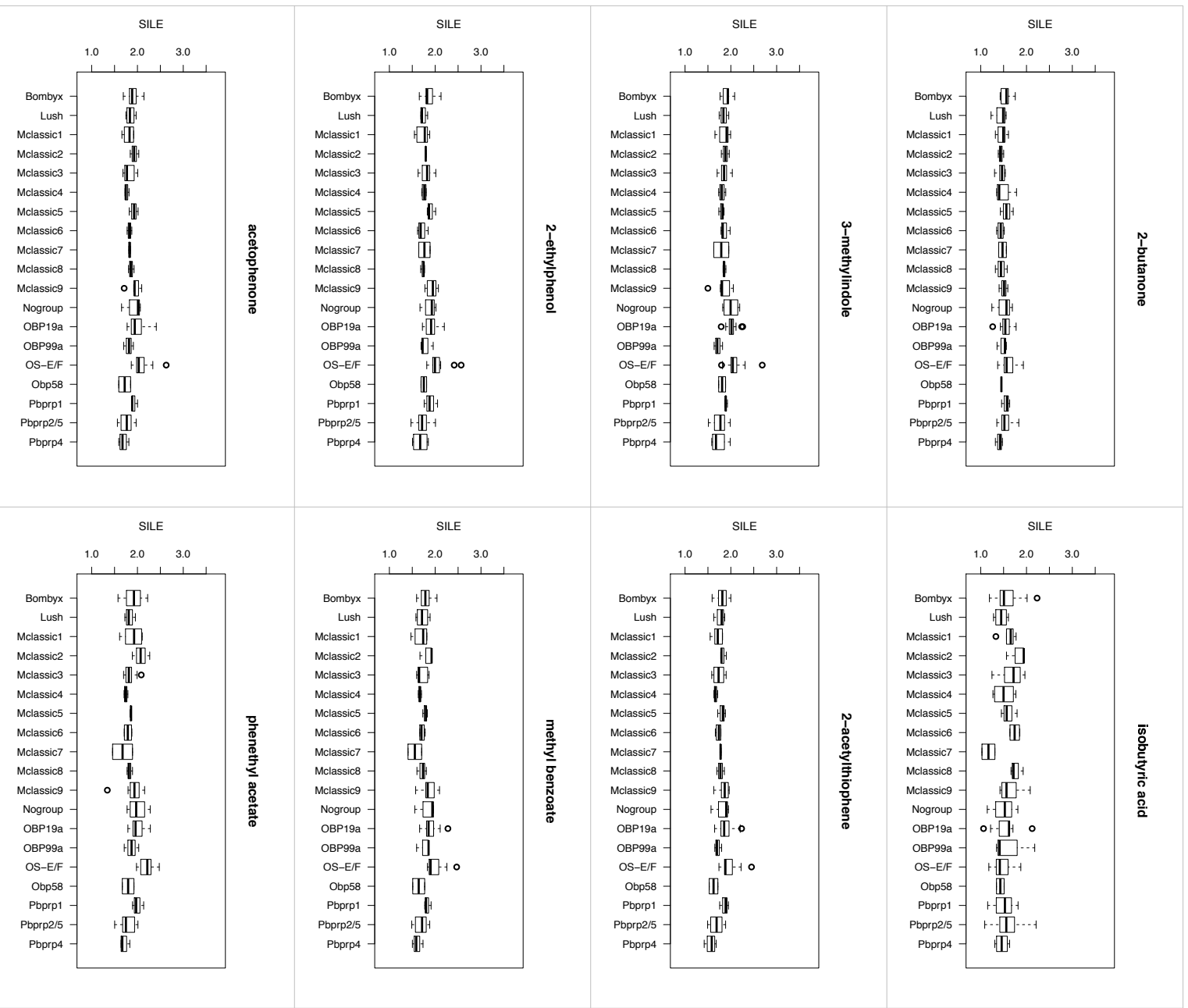
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



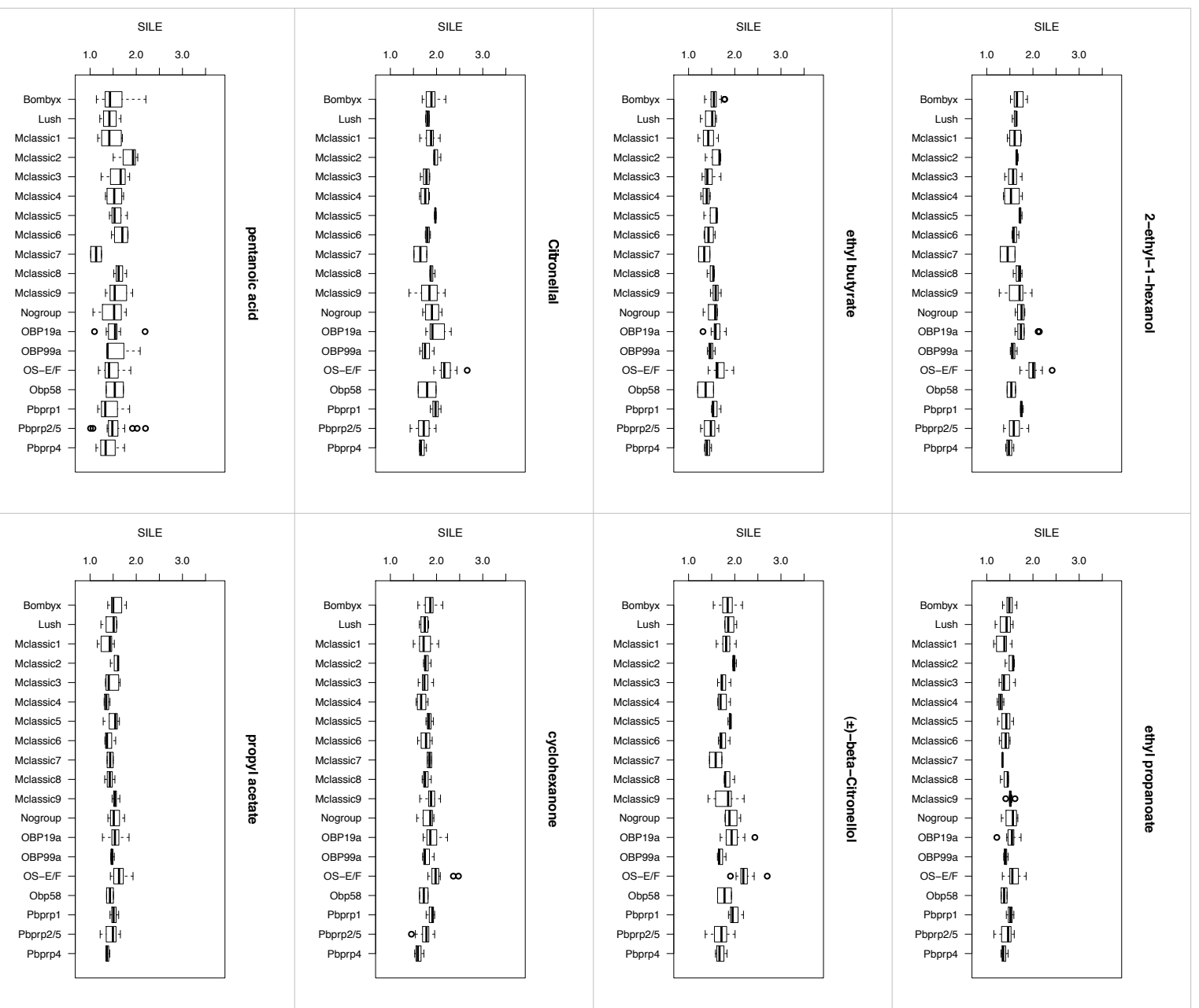
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



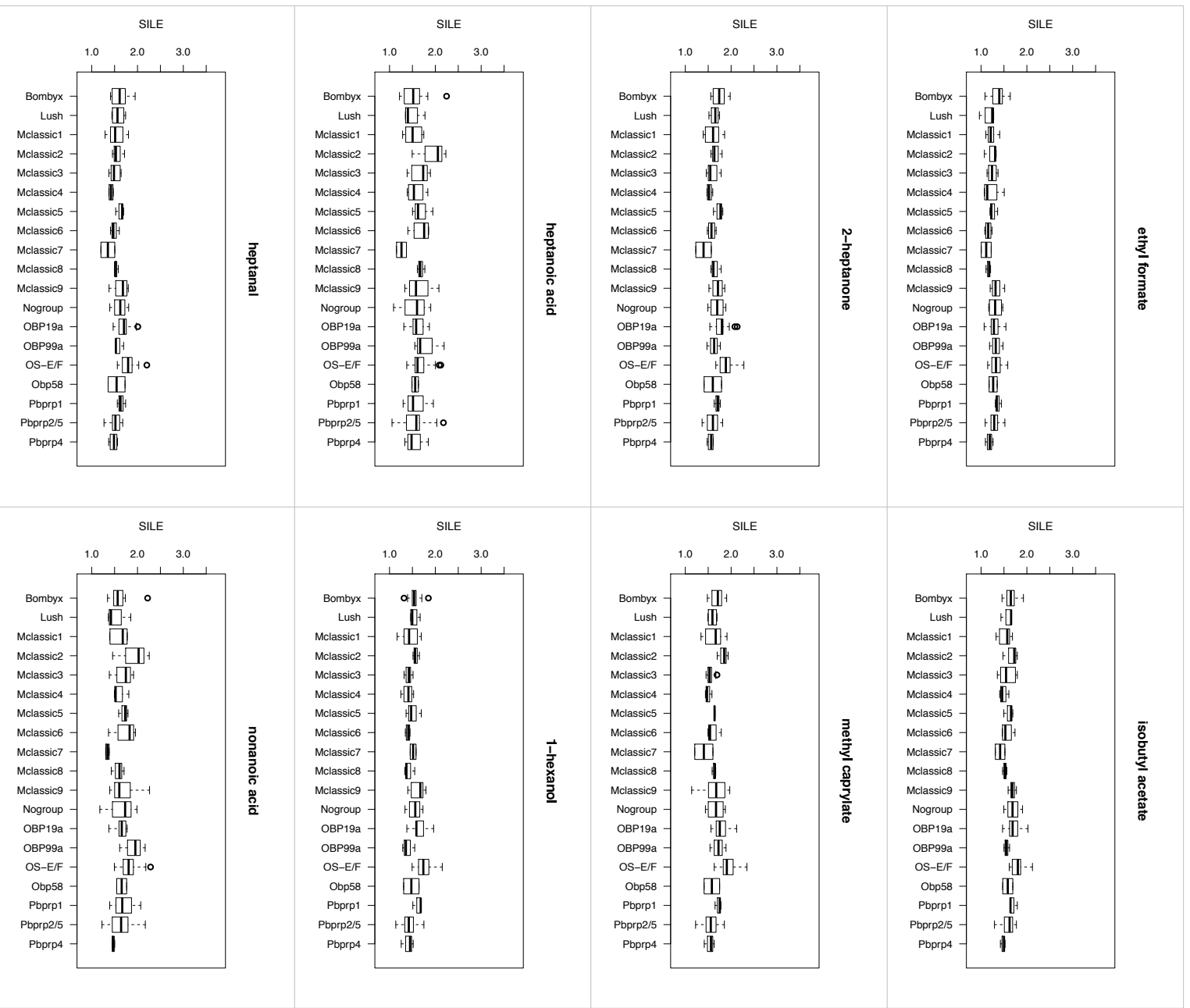
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



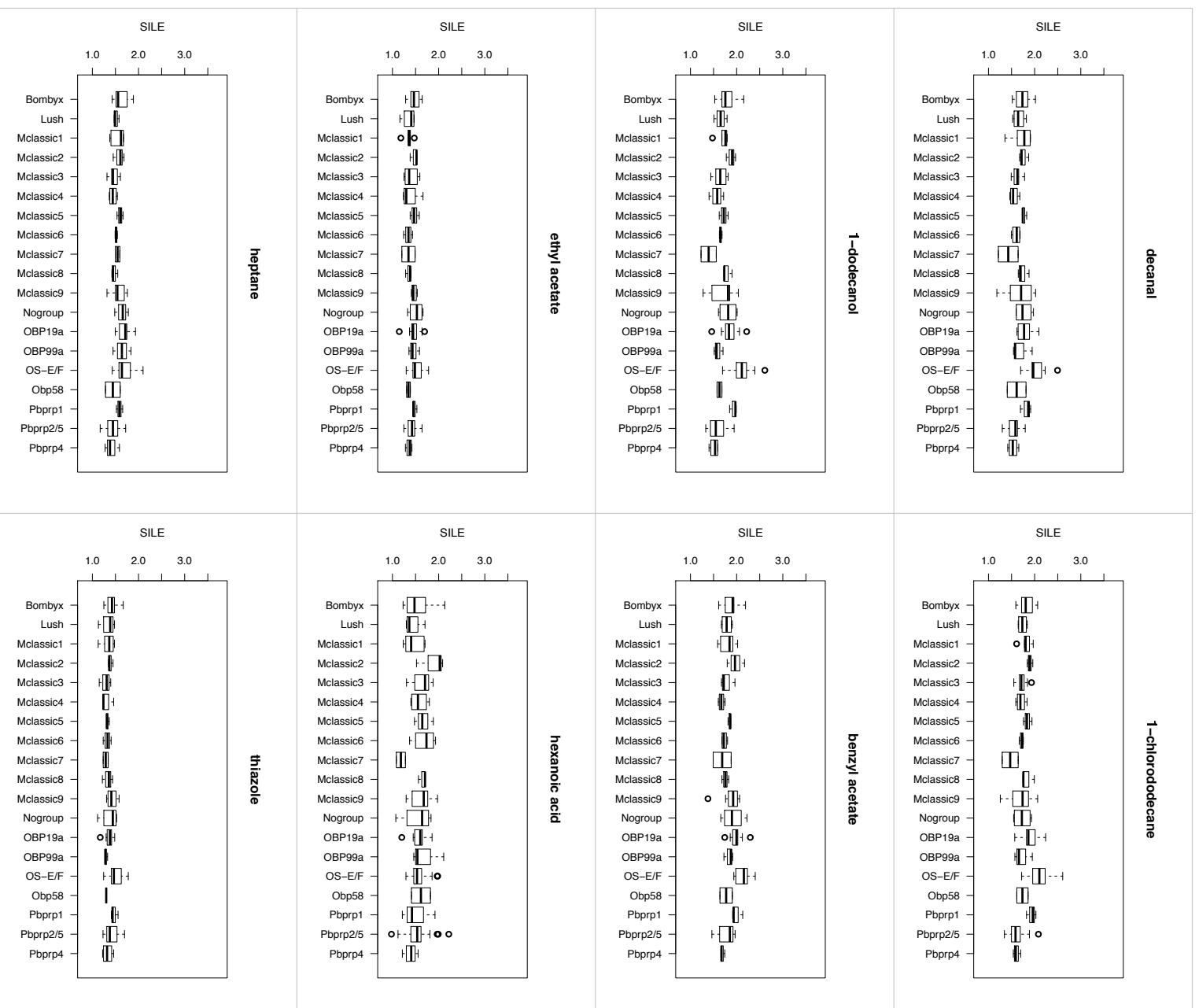
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



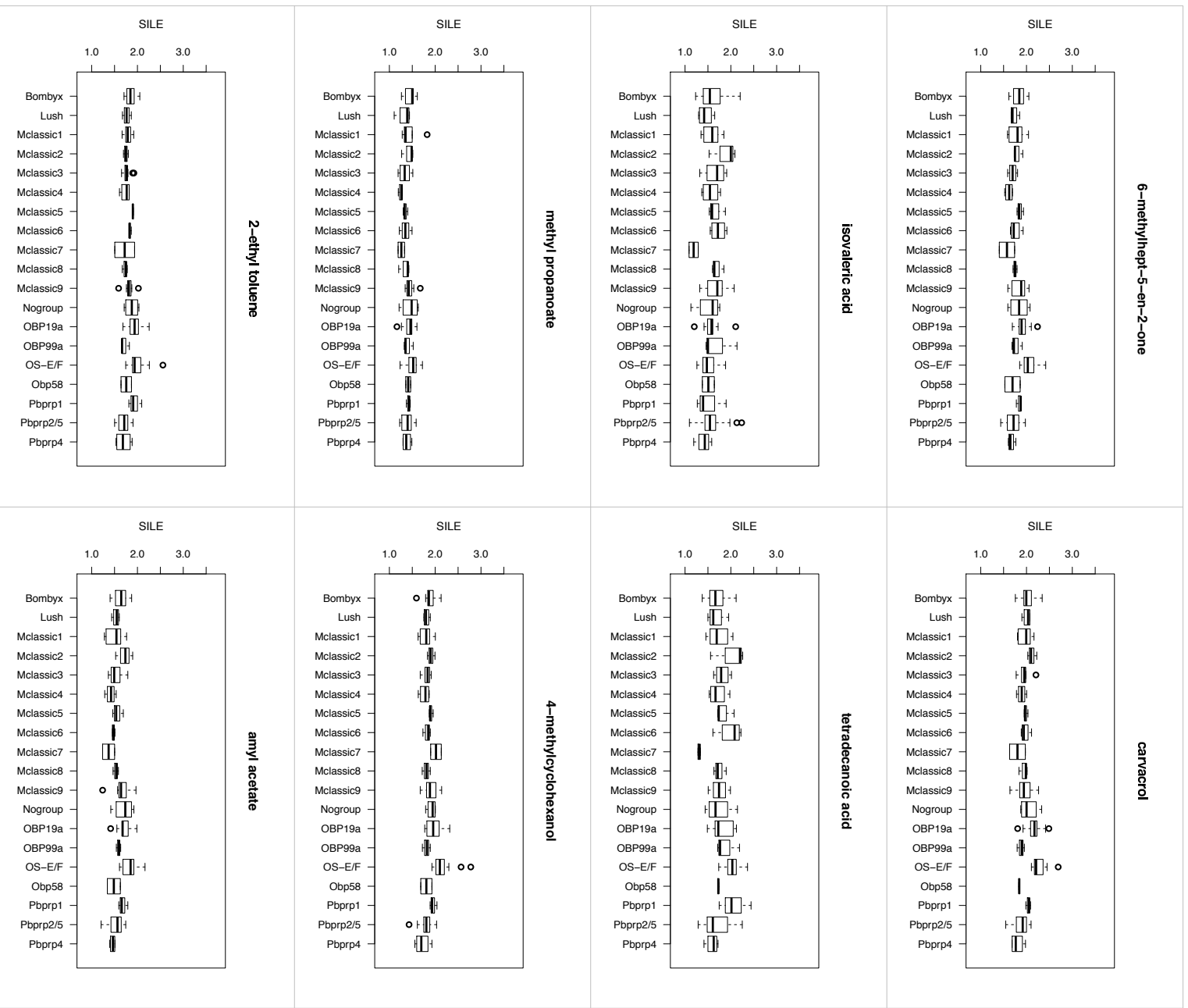
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



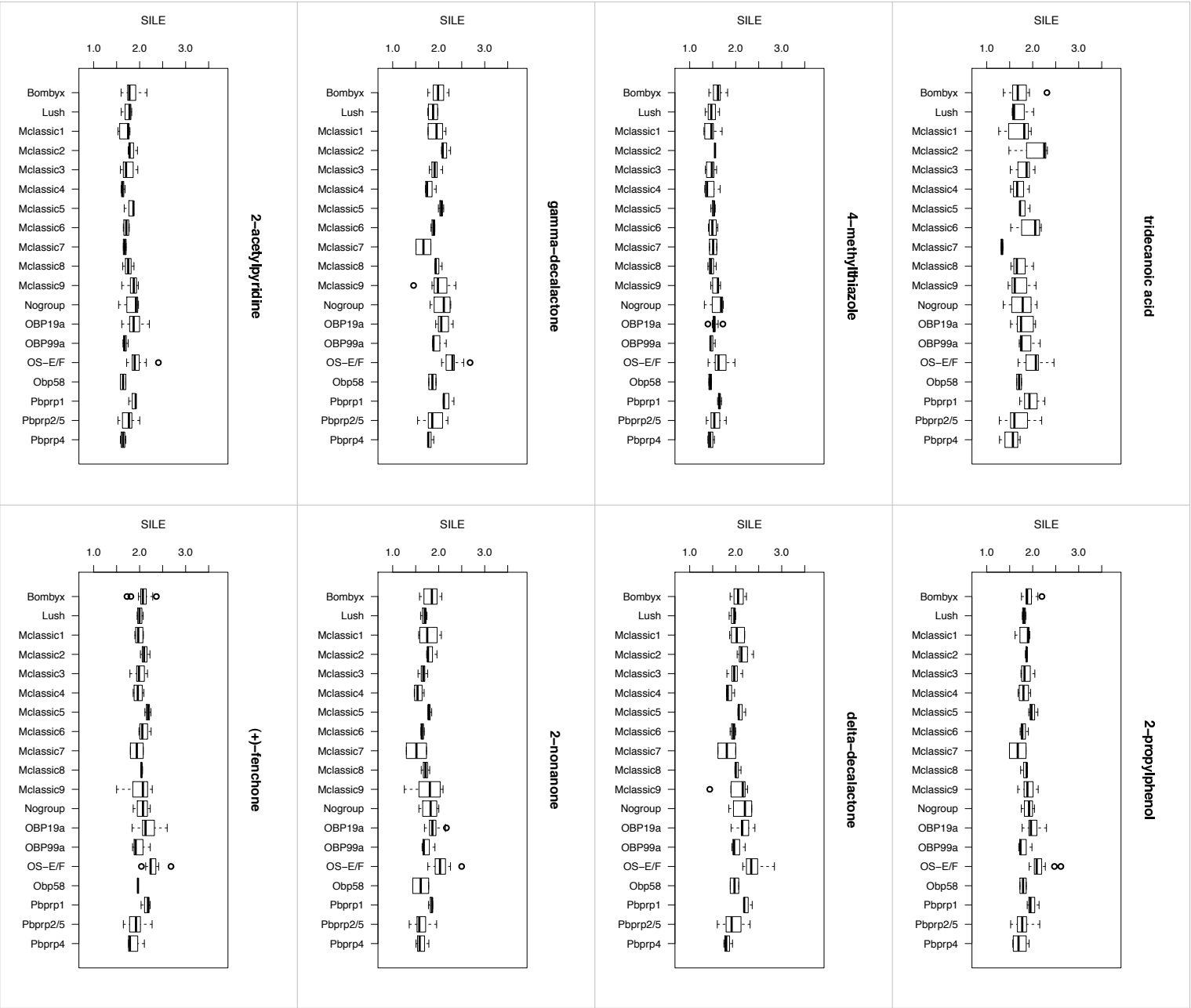
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



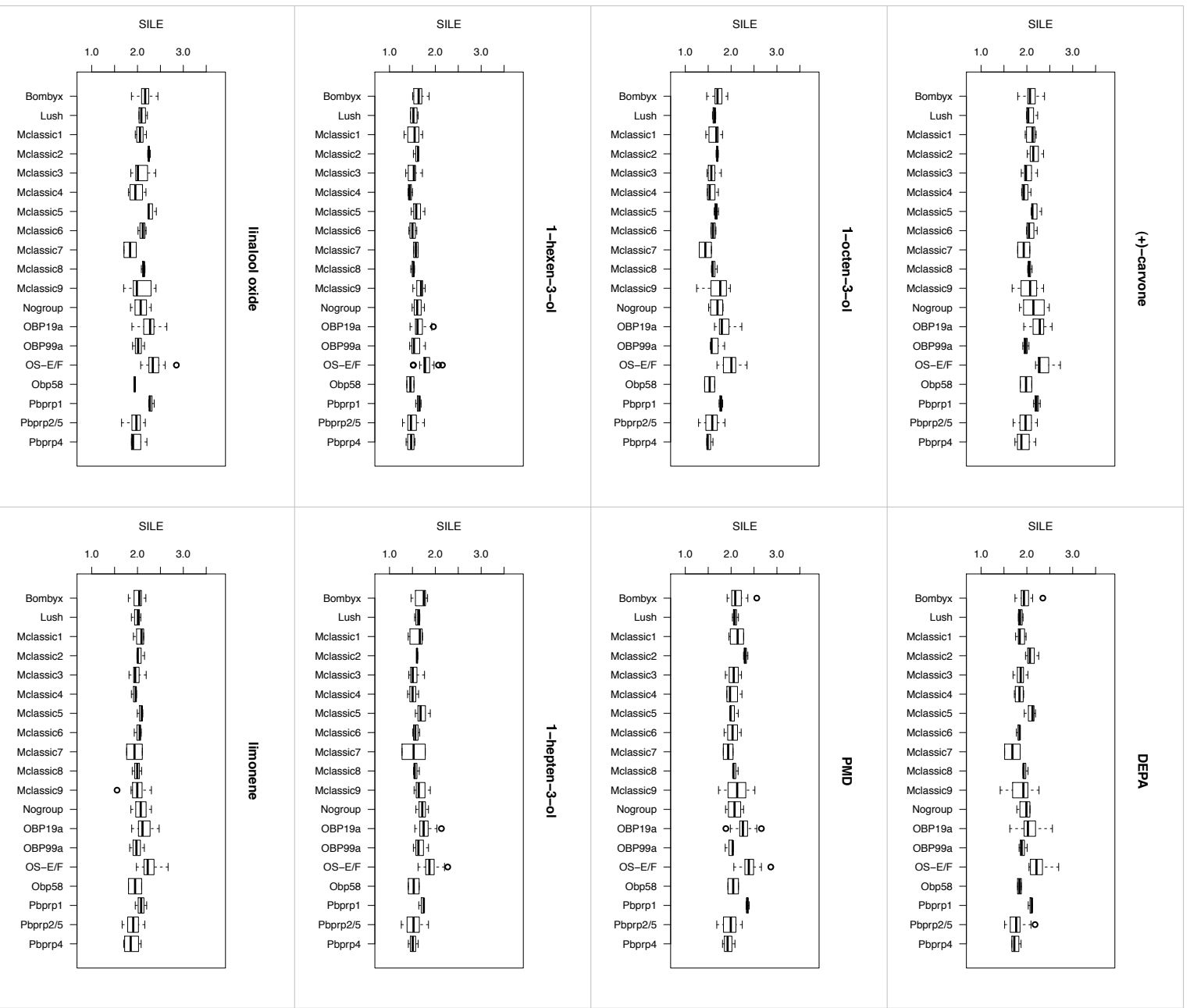
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



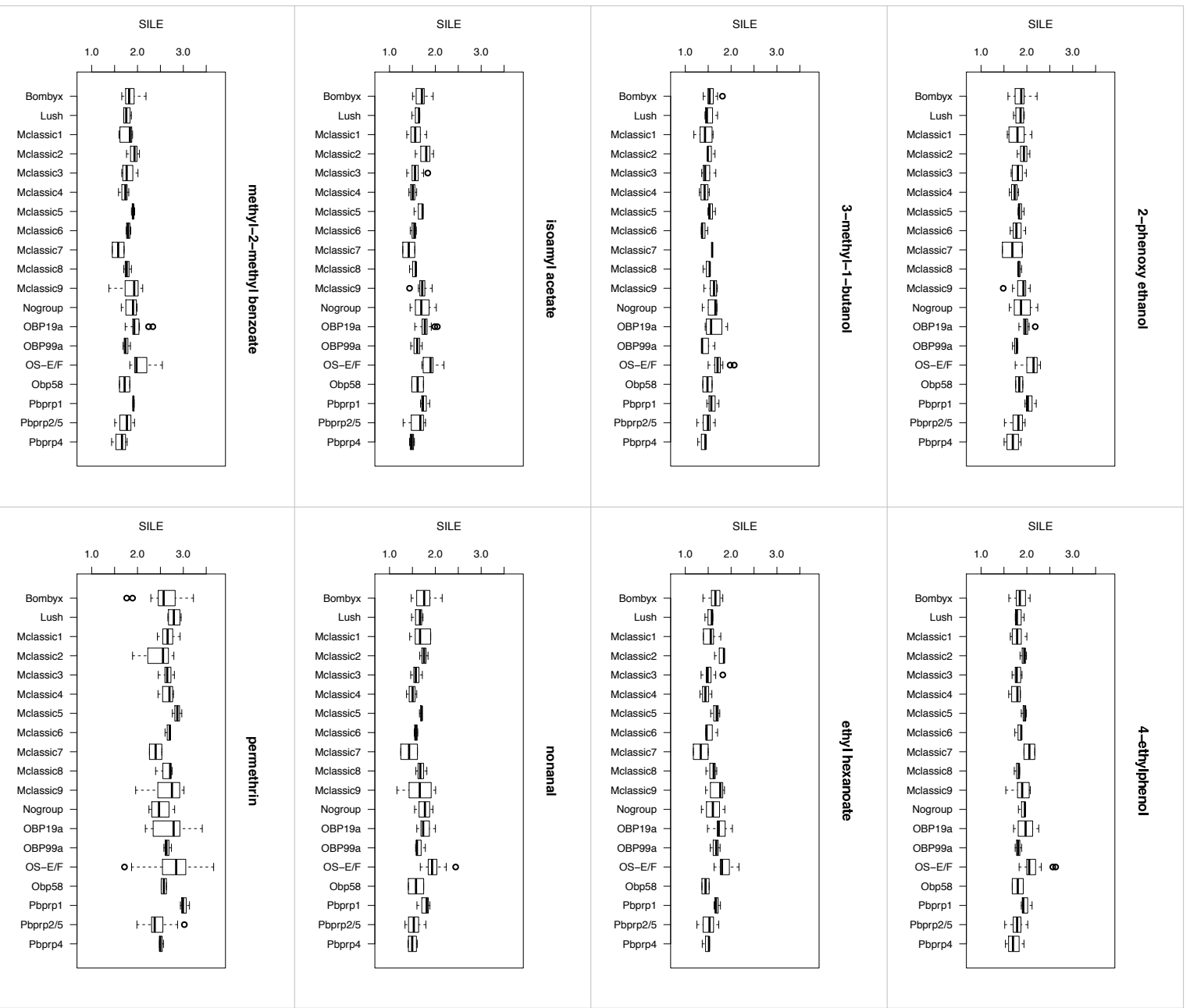
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



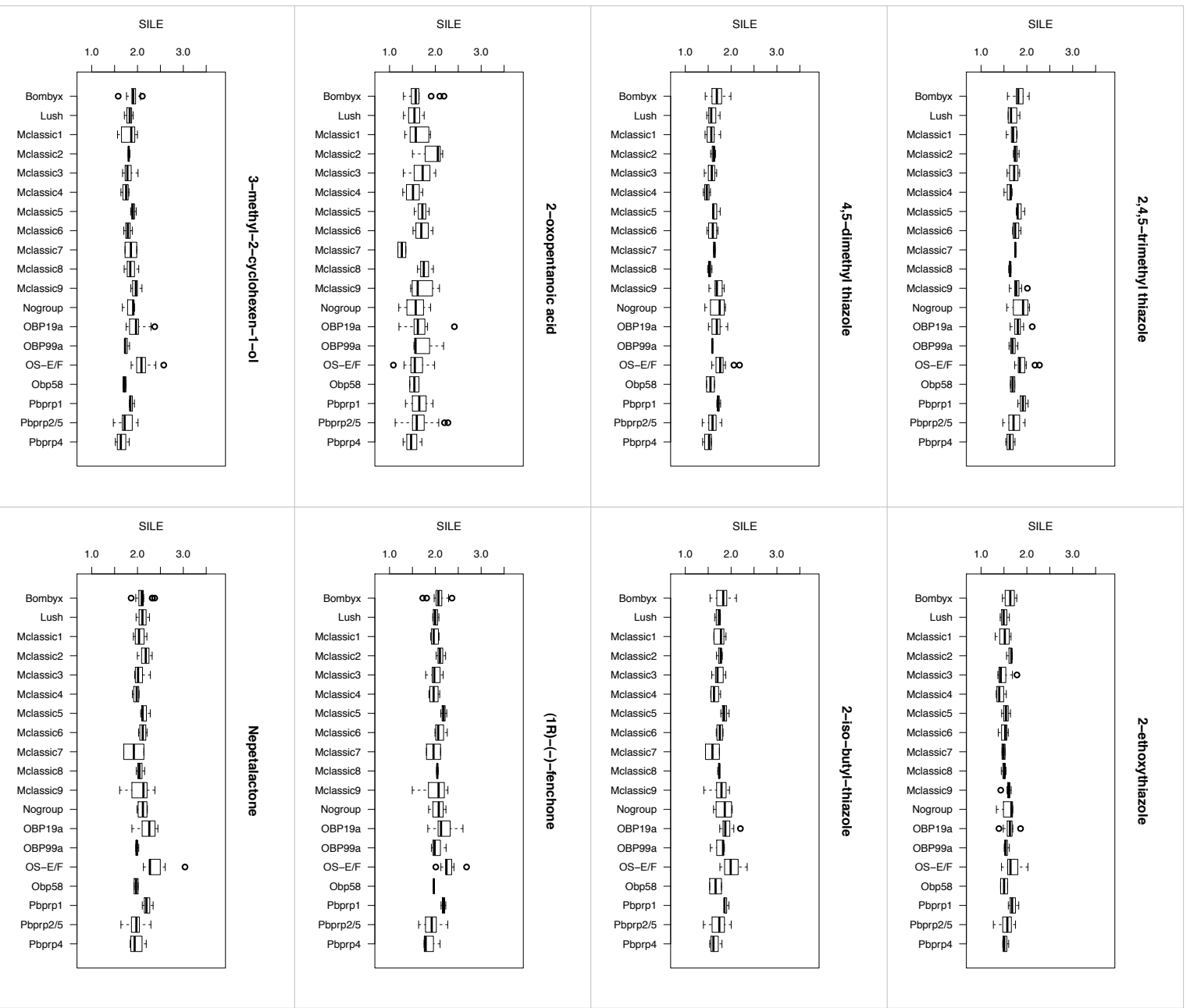
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



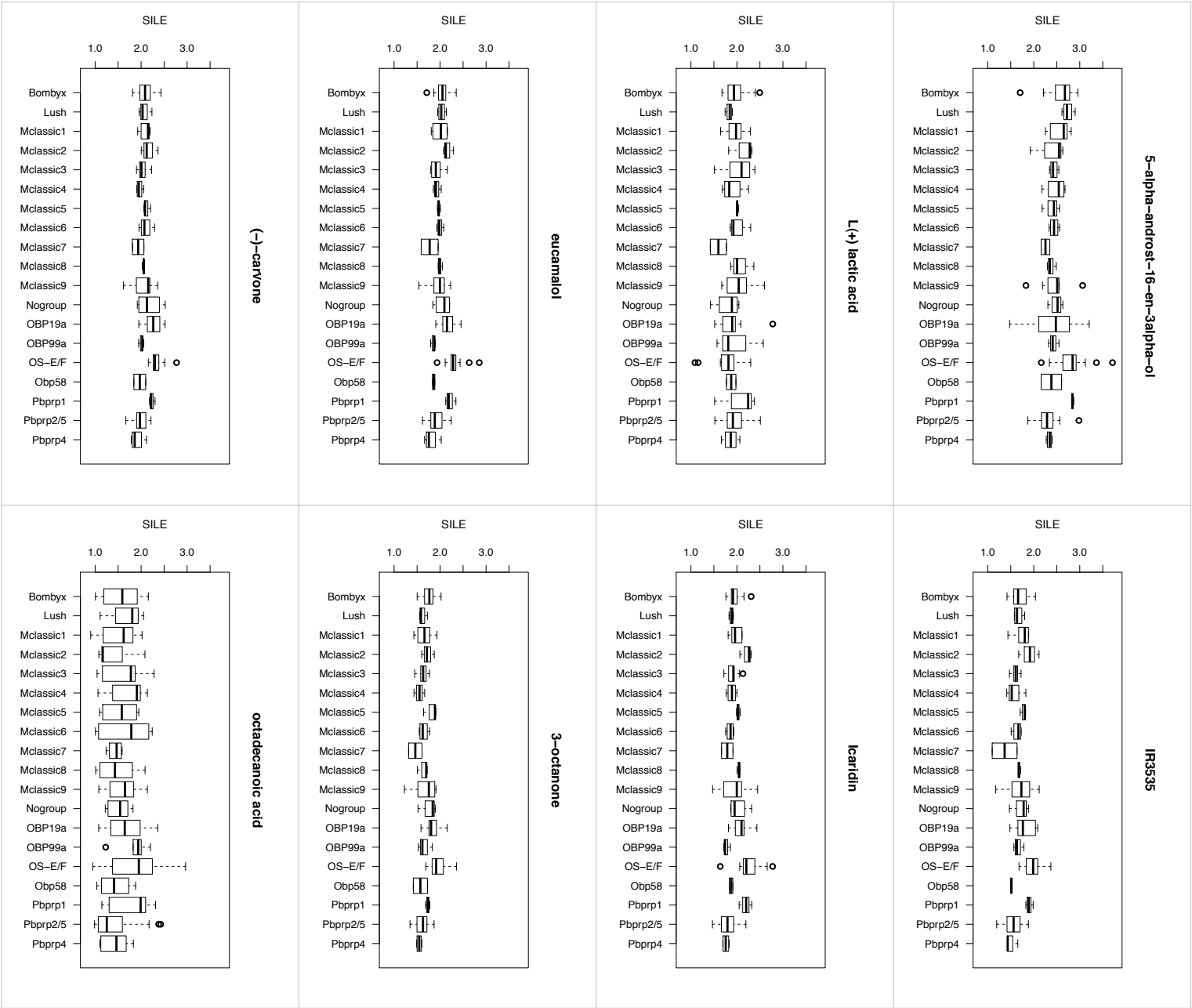
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



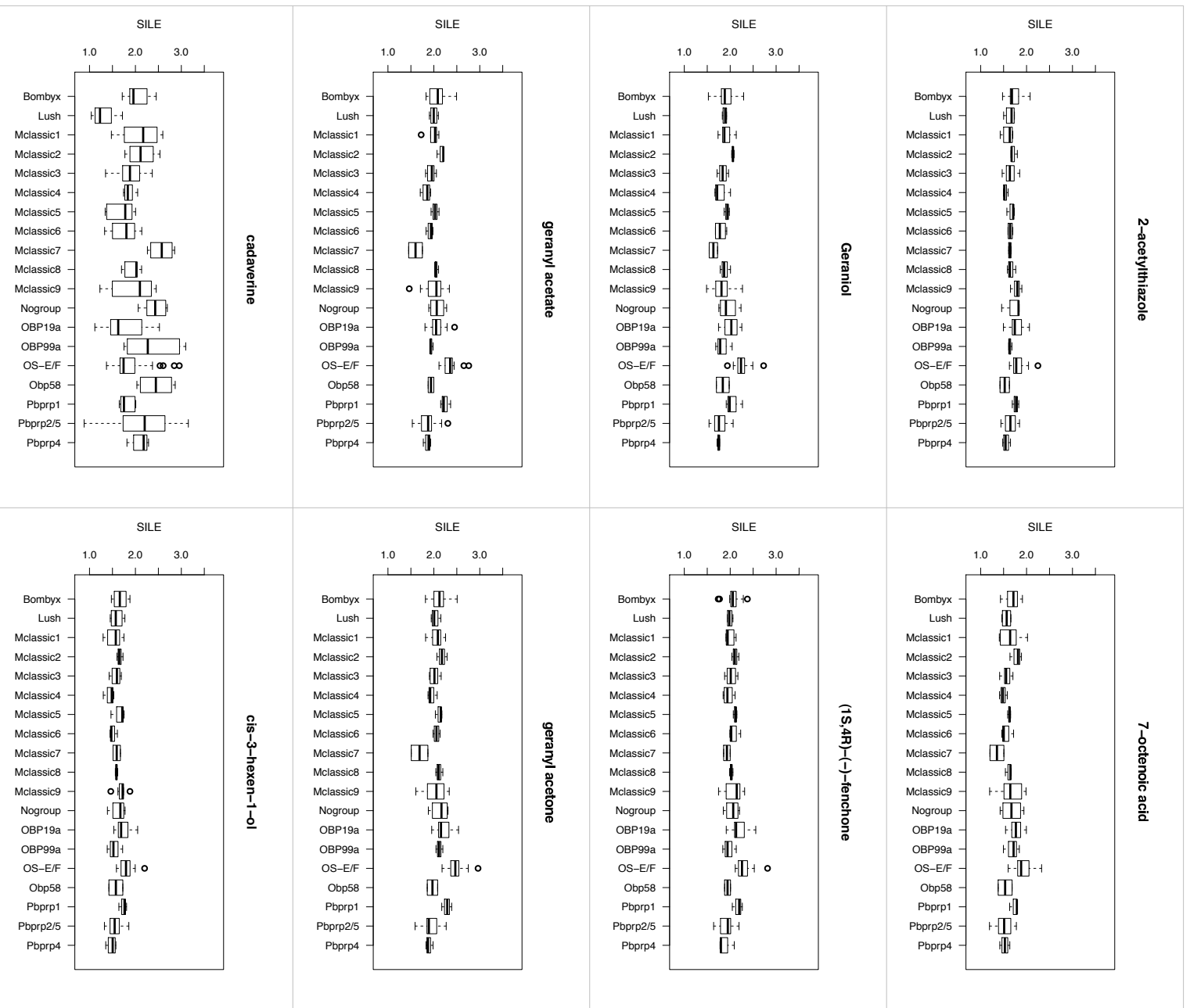
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



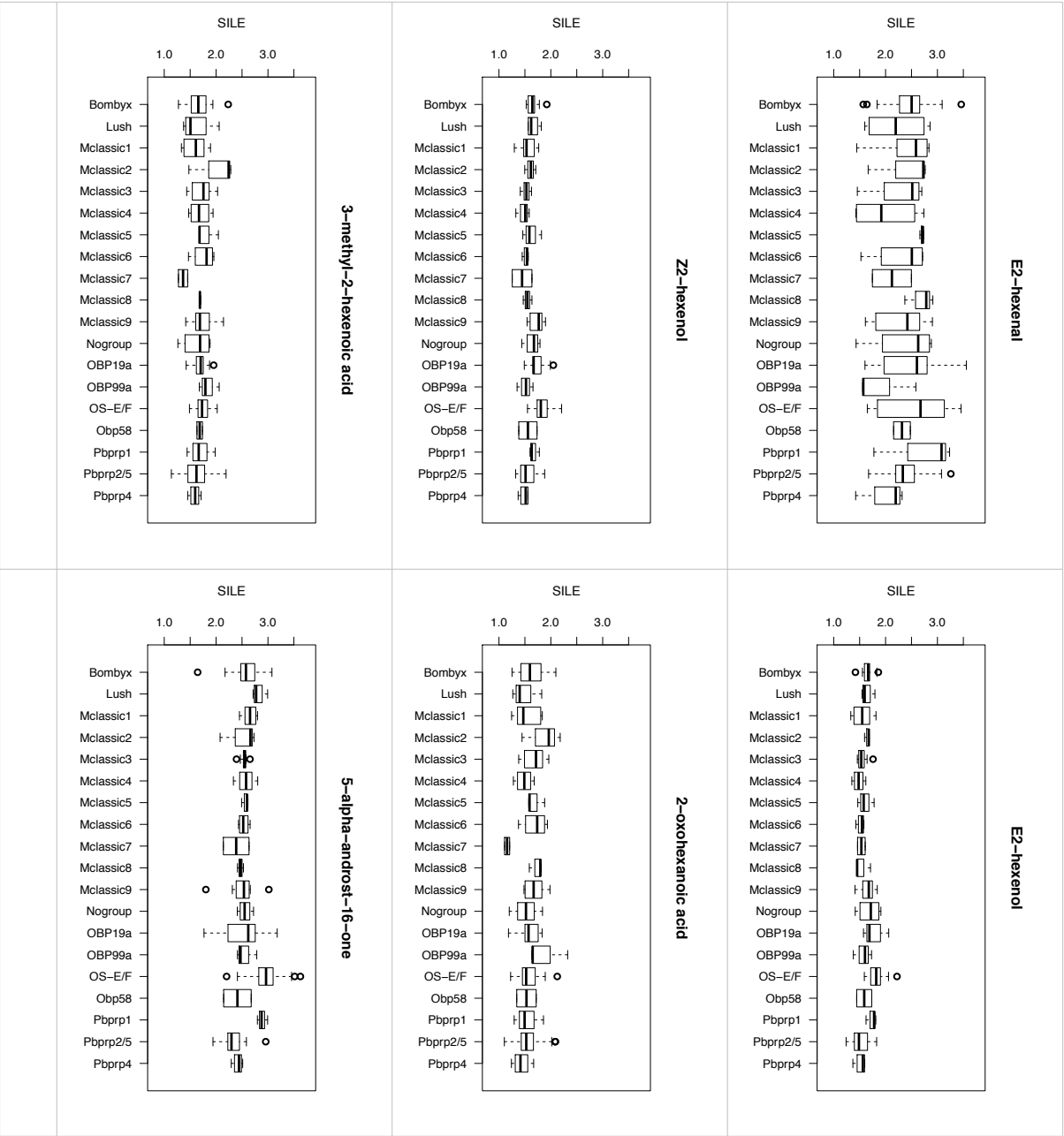
Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.



Supplementary Figure 3c. Analysis of ligands binding profiles towards the 18 established *Classic* OBP clusters. Shown are the boxplots of the size independent ligand efficiency (SILE) values for each of ligands towards every cluster. The plot represents the distribution of SILE values within each cluster of the classic OBPs with the median in the centre of each box represented as a thick black line. The entire plot stands as good representation for the comparison of the binding efficiency of a given ligand across the clusters.

Abstract

The role of odorant binding proteins in the olfaction of mosquitoes, the primary mechanism of human host recognition, has been an important focus of biological research in the field of infectious disease transmission by these insects. This thesis provides an in depth knowledge of these proteins in three mosquito species *Anopheles gambiae*, *Aedes aegypti* and *Culex quinquefasciatus*.

A large scale analysis on these genomes has been carried out towards the identification of the odorant binding proteins in the mosquito genomes. Identification of many new OBP members, in particular in the *Aedes aegypti* and *Culex quinquefasciatus* species, and an extensive phylogenetic analysis presenting a novel classification of the OBP subfamilies of these mosquito species has been proposed. This results further demonstrates the extraordinary multiplicity and diversity of the OBP gene repertoire in these three mosquito genomes and highlights the striking sequence features that are nevertheless highly conserved across all mosquito OBPs. Owing to the availability of homologous structures from mosquitoes or related species, the 3D structure modelling of all the *Classic* OBPs from the three genomes (representing in total 137 structures) has been performed. This was completed by large scale docking studies on these structures by screening a large set of compounds that are known to be mosquito attractants or repellents. These provide many exciting new insights into the structural and functional aspects towards understanding the efficacy of some repellents and of some attractants from human emanations. Through molecular dynamics simulation, the structural changes observed in an OBP bounded to an odorant when pH conditions are modified were characterized and the probable mechanism of ligand binding and release is presented. This work provides the first insights to many of the long awaited questions on the genomic, structural and functional characterization of mosquito OBPs and can be viewed as a reliable starting point for further experimental research focussed on these aspects.

Résumé

Dans le système olfactif des moustiques, les protéines liants les molécules odorantes ou *odorant binding proteins* (OBPs) interviennent dans les toutes premières étapes permettant d'aboutir à la reconnaissance de leurs hôtes et font l'objet d'un intérêt croissant dans les recherches sur la transmission des maladies infectieuses par ces insectes. Le travail présenté a pour objet d'approfondir les connaissances sur ces OBPs dans trois génomes de moustiques, tous vecteurs de maladies infectieuses : *Anopheles gambiae*, *Aedes aegypti* et *Culex quinquefasciatus*.

Une analyse à l'échelle de ces génomes a été réalisée et a permis d'identifier un nombre important de nouveaux gènes d'OBPs notamment chez les espèces de moustiques *Aedes aegypti* et *Culex quinquefasciatus*. Complétée par une étude phylogénétique du répertoire complet de ces gènes dans les trois génomes étudiés, cette analyse a permis d'établir une nouvelle classification des sous familles des OBPs. Ce résultat démontre l'extraordinaire multiplicité et diversité des gènes impliqués dans l'olfaction chez ces espèces de moustiques tout en mettant en lumière certaines propriétés des séquences des OBPs qui sont hautement conservés chez les moustiques.

Grâce à la disponibilité de certaines structures d'OBPs de moustiques ou d'autres insectes apparentées, des modèles structuraux de tous les OBPs de la sous famille dites *Classic* dans les trois génomes, soit au total 137 structures, ont été construits. Ces structures ont servi de base pour le criblage à grande échelle par docking moléculaire d'une chimiothèque de 126 molécules odorantes connues pour leurs propriétés attractives ou répulsives vis-à-vis des moustiques. Ces résultats fournissent pour la première fois, les bases structurales et fonctionnelles pour la compréhension au niveau moléculaire de l'efficacité de certains agents répulsifs tout comme de l'attractivité de certains agents provenant des émanations humaines. Par simulation de dynamique moléculaire, les changements qui s'opèrent dans une de ces OBPs lorsque celle-ci, liée à une molécule odorante, se retrouve dans des conditions de pH modifiée ont été caractérisée et un mécanisme probable par lequel ces OBPs participeraient à la reconnaissance et la libération des molécules odorantes est proposée. Cette thèse fournit des éléments de réponses importants quant à la caractérisation génomique, structurale et fonctionnelle des OBPs de moustiques et peut servir de base de départ pour des recherches expérimentales plus approfondies sur ces aspects.